Elementary Social Statistics

McGRAW-HILL PUBLICATIONS IN SOCIOLOGY

EDWARD BYRON REUTER, *Consulting Editor*

ELEMENTARY
SOCIAL
STATISTICS

By Thomas Carson McCormick
Professor of Sociology, University of Wisconsin

FIRST EDITION
FOURTH IMPRESSION

★ ★ ★

To My Wife

LILLIE GRIFFITH MCCORMICK

Preface

This beginning textbook in statistical methods has been written to meet the needs of undergraduate college students who are concentrating in sociology and related subjects. In the choice of methods, in the character of the illustrative data and problems, and in emphasis throughout, it differs from the texts in economic or educational statistics that have generally been used by such students.

The chief purpose has been to provide students who expect to become professional sociologists with the necessary groundwork for more advanced training in quantitative research methods. Familiarity with the topics included, however, should enable those who take no further courses in statistics to understand most of the statistical studies and references that now appear in the sociological journals and literature. Nonprofessional students who go through the course should learn to appreciate some of the difficulties involved in the study of social problems, and to be more wary of careless and prejudiced thinking in this field; for mathematical statistics represents a rigorous form of applied logic.

Unfortunately, most students who elect to specialize in sociology have no mathematical training beyond high school algebra. This fact has compelled the omission of mathematical derivations, with the exception of a few very simple ones. As a substitute, an attempt has been made to point out assumptions that should be watched in using the various formulas. Students who plan to go on in the subject, however, should begin at once to build up an adequate mathematical background.

Because of its complications and as yet very infrequent use in sociological research, small-sampling theory has for the most part been omitted from this elementary treatment.

The amount of material covered is more than enough for a semester's work with an average class, so that some selection of topics is possible for the instructor. Under certain circum-

stances, it may be advisable to omit the less easy sections of chapters IX, XI, XII, XIII, and XIV.

Constant practice in working statistical problems is indispensable for mastery of the subject. The problems given at the end of each chapter are intended to be only suggestive; they should be greatly multiplied for laboratory purposes.

Thanks are due Professor E. A. Gaumnitz of the University of Wisconsin, who has read the manuscript and made helpful suggestions, and Mr. Robert J. Hader, who has eliminated numerous minor errors.

Special acknowledgment is made of permission by Prof. R. A. Fisher and his publishers, Oliver & Boyd, Edinburgh, to use the Table of Chi-square and the Table of Values of the Correlation Coefficient for Different Levels of Significance, which appear as Tables 2 and 4 in the Appendix of this book. Many other publishers have been kind enough to grant permission to use tables and material, specific acknowledgment of which has been made in place.

Thomas C. McCormick.

Madison, Wis.,
August, 1941.

Contents

CHAPTER X

CHAPTER XI

CHAPTER XII

CHAPTER XIII

CHAPTER XIV

PART I

Statistics in Social Research

CHAPTER I

INTRODUCTORY

1. The Origins of Statistics.—The word *statistics* was used in Great Britain and *Statistik* in Germany as early as the eighteenth century to refer to collections of information of any kind about a state (*state*-istics). As time passed, "statistics" came to be limited to quantitative data or figures on wealth, taxes, marriages, baptisms, deaths, and the like. Distinguished pioneers in the field were the Germans, Achenwall and Büsching. Modern agencies representing this type of statistics are the census bureaus of the United States and other nations.

Mathematical statistics, a branch of mathematical theory, originated in investigations of birth and death rates and in efforts to solve problems growing out of games of chance. Among the great early vital statisticians were Graunt and Petty of England and Süssmilch of Germany. The fundamentals of the theory of probability were developed from the seventeenth to the nineteenth century by such eminent mathematicians as Pascal, Bernoulli, de Moivre, Laplace, and Gauss.

Elementary mathematical statistics was popularized in the nineteenth century by the Belgian, Quetelet, who applied it to a wide variety of topics, including physical anthropology and crime. He is sometimes called the father of *social statistics* as the extension of statistics to sociological problems may be termed.

A rapid expansion in mathematical statistics and its use in science occurred in England during the first quarter of the present century through the work of Karl Pearson, following earlier efforts by Sir Francis Galton. These two men were biologists, and Pearson was a mathematician as well. As a result of this phase, modern statistical methods bear the imprint of adaptation to biological data.

Mathematical statistics has gradually become a major method of research in the fields of agriculture, biology, educational

psychology, psychology, geography, and physical anthropology. Among the social sciences, education, economics, and social psychology at present lead in the proportion of statistical studies published, with sociology fourth and political science fifth. Statistical analysis is still rare in cultural anthropology and history. In a different direction, statistics is finding application in mathematical physics, engineering, and medicine.

In this book, we shall be interested in elementary statistical methods only as tools of investigation in sociology and related social sciences.

2. Quality and Quantity.—A qualitative difference implies a difference in nature, such as we recognize between a family and a church. A quantitative difference refers to a variation in amount between two or more instances of the same quality: for example, an intelligence quotient (I.Q.) of 112 is 14 units greater than an intelligence quotient of 98.

Different qualities must be compared in terms of common subqualities (common denominators), as a Presbyterian family and a Methodist church, a family of five members and a church of 500 members. A pure quality, moreover, can vary only in amount. It follows that all comparison must consist in noting what qualities are and are not common to *A* and *B*, and how each common quality varies in amount from *A* to *B*. The city and the country may be compared in terms of common qualities like population density, birth rate, death rate, incidence of tuberculosis, intelligence quotients, honesty, and so on; but in each instance the difference must be in terms of amount. Thus the birth rate of the city is 15 per 1,000, that of the country is 22 per 1,000; and country people are believed to be *more* honest than city people. The last judgment is no less quantitative in nature because it is impressionistic and rough.

Because comparison is basic to knowledge, and quantitative judgments are inseparable from comparison, quantitative judgments are unavoidable in science. It is thus easy to understand why scientists have gradually developed more and more systematic and reliable ways of making quantitative judgments, such as we have in the many branches of mathematics, including mathematical statistics.

3. Statistics, the Method of Probabilities.—The questions in which social scientists are interested do not have exact or certain

quantitative answers. For example, if we ask what is the relation between the occurrence of divorce and the presence of children in the home, we find that divorce occurs both among couples with children and among couples without children, but relatively more often in the case of the latter. We cannot say that divorce takes place only when there are no children; but we can say that divorce is reported in so many childless marriages per 1,000, and in so many fertile marriages per 1,000. Or, expressing it a little differently, we can say that the chances of divorce are X in 1,000 in the case of a childless couple, and Y in 1,000 in the case of a fertile couple.

Statistical methods are specially designed for the analysis of quantitative[1] data like those above that result from many causes, some or all of which cannot be completely controlled. Outside the scientific laboratory, and even in much laboratory research, adequate control over all factors is out of the question. For this reason, the statistical method has general application.

Mathematical statistics is a direct logical extension to practical situations of the exact quantitative methods used in the laboratory experiments of the physical sciences. When precise measurement and complete control over all factors are possible, a mathematical equation can be set up from which the value of a dependent factor, Y, can be estimated exactly for any given value of an independent factor, X. For example, if we know the distance, X, of an object from the ground, we can calculate from the law of falling bodies the time, Y, it will take for the object to fall *in a vacuum*. When we actually drop an object under these controlled conditions, a stop watch will always register the length of time predicted by the equation. The likelihood that the period observed in any competent repetition of the experiment will be that computed from the equation is certainty.

If, however, the object happens to be a feather which is dropped under ordinary atmospheric conditions rather than *in a vacuum*, the situation is different. In proportion as the factors are uncontrolled or unknown, the stop watch will no longer register the time predicted by the law of falling bodies. Nevertheless, if a large number of experiments are made by

[1] By quantitative data are meant data that can be measured or counted, as discussed below in Chap. II.

dropping the feather under ordinary atmospheric conditions from the same distance, the time of falling will be found to vary around some average time, being sometimes more and sometimes less.   Similarly, the average time of falling from other distances, X, can be found, and an equation worked out from which the average time of falling, Y, can be estimated for any distance, X. Then by studying the varying time required for the object to fall a given distance, it may be established that in say two-thirds of the trials the time does not vary from the average time, say 2 sec., by more than say 0.1 sec.   This enables us to make a prediction from our empirical equation.   We can say that if our feather is dropped under ordinary atmospheric conditions from a given height, the time required to fall will, two out of three times, in the long run, vary from an estimated average of 2 sec. by not more than 0.1 sec. in either direction.   That is, in two out of three trials, the time of falling will be between 1.9 and 2.1 sec.

This is, broadly speaking, the kind of estimate that mathematical statistics furnishes in the social sciences.   In essence it is always a calculation of probabilities.   The "pure" mathematical formula of the laboratory is merely a special case of the statistical equation, being the limit that the latter approaches as the amount of control and precision of measurement are increased.   If sociological data could be exactly controlled and measured, the element of probability would disappear, and the statistical equation would become a precise one like the law of falling bodies.

4. Representative Data.—Most sociological studies, statistical or otherwise, deal with samples rather than with complete data. If farm life in a given state is to be investigated, certain farms are taken as a sample to represent all the farms in the state regarded as the *universe*.   The essential requirements of a good sample are that every item in the universe from which the sample is drawn shall have an equal chance of being included in the sample, and the sample must be large enough to include every kind of item in the universe in something like the correct proportions.   The proper size of the sample depends somewhat on how much the items in the universe vary among themselves.   Poor samples that include items from outside the universe they are intended to represent, that omit important elements of the

universe, or that include elements of the universe in the wrong proportions, are a fertile source of false conclusions in social research. A large part of mathematical statistics deals with the problems of sampling.

5. Statistics and the Individual.—It is commonly thought that statistics cannot deal with the individual, but must confine itself to group averages. There is really nothing to prevent a statistical investigation of an individual. An individual may be readily analyzed into factors or units of various kinds, and the relationships of these to other factors in the same personality and in the environment can be studied by the same methods that are now used in studying groups of individuals. As a matter of economy, however, society will seldom want to subject individuals to scientific study except as types, which, of course, lead back to group averages.

6. Interpretation of Statistical Results.—Statistics employs figures and mathematical symbols that represent definite factors in a particular problem. In interpreting statistical results, therefore, care must be taken that each symbol is given the same meaning that was assigned to it at the beginning of the problem, and to which no important exceptions were allowed during the study.

It is sometimes puzzling to understand the reasons for a statistical fact, and offhand explanations may be found at the end of even careful studies. But if the original study was not sufficiently inclusive to clarify some point of interest, its reliable explanation can consistently come only from further research. For example, if an investigation discovers that a larger proportion of women are married in cities where the number of men exceeds the number of women than in cities where the two sexes are equal in number, or where women outnumber men, one may speculate that this is because men do the proposing. It should be made clear, however, that such an explanation is only a plausible "hunch," which should be tested if it is considered of enough importance.

Difficulty may also be experienced in interpreting just what certain statistical concepts mean, *e.g.*, correlation coefficients, averages, or tests of statistical significance. The only help here is a clearer understanding of statistical methods, and especially of the mathematical assumptions that underlie them.

7. Statistics Not a Mechanical Method.—Although the statistical method allows data to be treated by systematic and standardized techniques, it is a serious mistake to suppose that it is a mechanical method that may be substituted for hard and original thinking. On the contrary, mathematical statistics is merely a set of powerful logical tools that call for a high type of judgment and skill for their successful use. The statistical investigator must know what techniques are valid and effective for a given problem, and when quantitative methods are not appropriate at all. He needs insight to select worthwhile problems, and intimate knowledge of the data to interpret his findings, no less than does any other type of investigator.

8. Simplicity the Ideal.—The experienced statistician always prefers simple to complex methods, when the two are equally effective. The beginner will do well not to yield to the temptation to depart from this sensible rule.

Exercises

1. Briefly summarize the history of statistics and the extent of its use as a method of research.

2. Distinguish between quality and quantity. Illustrate.

3. Can you find an exception to the proposition that all comparison is quantitative?

4. To what general kind of research situation is statistics appropriate, and why? Illustrate.

5. What is the relationship of the statistical equation to the mathematical "law" of physics?

6. *a.* How exactly can predictions be made by means of statistical methods?

b. How serious a handicap does this impose on social research?

7. What is the likelihood in the field of social research that the statistical method will some day be replaced by exact mathematical formulas like those of physics? Explain.

8. Comment briefly on the following published statement:

"*Jobless Survey to Bare Truth, C. C. Head Says:* Pres. George Davis of the Chamber of Commerce of the United States said Saturday an impartial survey of the employable jobless would show their numbers had been exaggerated and disprove alleged needs for spreading work by reducing working hours.

"He said the chamber recently employed a statistical agency to make a sample survey of 100 relief recipients in a representative city of more than 100,000 population. The names of 50 men and 50 women

were picked at random from Federal and local governmental relief rolls in the city.

"The survey showed, he said, that 44 out of the 100 never had been employed in private business. Seventeen were over 70 years *and* 82 *never had a bank or savings account.*

"He says the figures point out that the greater number of those labeled as unemployed could not or would not work in private industry even if jobs were available."

9. Give an example of representative and unrepresentative, adequate and inadequate sampling that might occur, or has occurred, in social research.

References

BERNARD, L. L., ed.: *The Fields and Methods of Sociology*, Farrar & Rinehart, Inc., New York, 1934.

DAMPIER-WHETHAM, W. C. D.: *A History of Science*, Chap. X, The Macmillan Company, New York, 1930.

GIDDINGS, F. H.: *Studies in the Theory of Human Society*, Chap. VI, The Macmillan Company, New York, 1926.

LUNDBERG, G. A.: *Foundations of Sociology*, The Macmillan Company, New York, 1939.

MILLS, F. C.: On Measurement in Economics, in the *Trend of Economics*, Rexford Tugwell, ed., Alfred A. Knopf, Inc., New York, 1924.

OGBURN, W. F.: Sociology and Statistics, Chap. XXX, in *The Social Sciences*, W. F. Ogburn, and Alexander Goldenweiser, eds., Houghton Mifflin Company, Boston, 1927.

RICE, STUART A., ed.: *Methods in Social Science: A Case Book*, University of Chicago Press, Chicago, 1931.

SMITH, JAMES G.: *Elementary Statistics*, Chaps. XIX and XXV, Henry Holt and Company, Inc., New York, 1934.

WALKER, HELEN M.: *Studies in the History of Statistical Method*, The Williams & Wilkins Company, Baltimore, 1931.

CHAPTER II

THE QUANTIFICATION OF SOCIAL DATA

1. Definition and Counting.—The methods of statistics are applicable only to data that can be expressed in some kind of countable units. Any event or quality that can be recognized can be counted. If we know a happy marriage when we see one, we can count the happy marriages in a sample of marriages. Nothing simpler can be done to a concept than to count how many times instances of it occur. If the concept is not sufficiently recognizable for its instances to be counted, one may fairly assume that it is not yet ready for any kind of scientific manipulation, except attempts to arrive at a more reliable definition.

2. Classification.—If a concept (*e.g.*, "conflict behavior") can be broken down into two or more subcategories (*e.g.*, "war," "revolution," etc.) that can be defined well enough to be told apart, its cases can be *classified*. Classification makes possible the counting of instances in each class, which may then serve as a basis for considerable statistical analysis. We have simple classification whenever data are sorted into categories that are entirely *unordered* with respect to amount. For example, we may classify our acquaintances as religious and nonreligious; we may classify Americans as native white of native parentage, native white of foreign parentage, foreign born, and so on. Data may also be classified with respect to two or more criteria at a time, as married couples by occupation of husband, by income, and by number of children. The points to watch in classification are careful, objective definition of the several categories in terms of criteria that can be recognized in the instances to be classified, and independent reclassification of the instances by other competent investigators, to determine the reliability of the classification. Logically, any classification should be based on the same criterion throughout. Thus, it would not do to classify some of the foreign born as Catholics

10

or Protestants, and the rest as Italians, Jews, Germans, and so on. Also, any classification should be totally inclusive of the class defined and exclusive of all other classes.[1]   That is, if we are dealing with all the foreign born in the United States, the Rumanians should not be omitted, nor should the American Indians be included.

3. Measurement of Amount.—The fact that any quality, such as happiness in marriage, varies in degree, sooner or later forces the sociologist to go beyond the mere counting of instances, and to attempt to *measure* the intensity of the quality in a given instance or set of instances.   For example, we may score the answers of married couples to a questionnaire and may regard the score of any couple as an index of the amount of happiness that they derive from their relationship.   The central problem is, again, to find a unit in terms of which at least the relative amount of the quality can be measured.   This is seldom easy to do, and must usually be approached through the devices of ranking, rating, or scoring.

4. Ranking.—Ranking, or the arrangement of the instances of a quality in order of amount, has been called the most elementary form of measurement.   We consider person A more cooperative than person B, B more cooperative than C, and so on. To increase the reliability of these judgments, the ranking may be done independently by several qualified judges, and the average ranks taken.   Greater accuracy is sometimes obtained by ranking each item with respect to every other item, *i.e.*, by all possible pairs.   Where qualified and careful judges cannot be obtained, ranking should not be used.   As soon as the instances of a quality are ranked, they become capable of a fair amount of statistical treatment, including rank correlation.[2]

5. Rating.—Similar to ranking is rating, or the classification of items into ascending, or *ordered*, classes.   There are usually three to seven of these classes.   An odd number allows for a median class, which is desirable.   Thus psychiatrists may rate persons in terms of their intelligence as Mentally Defective, Slow-dull, Slow, Average, Fairly Intelligent, Distinctly Capable,

---

[1] See "classification" in any text in logic, *e.g.*, E. A. Burtt, *Principles and Problems of Right Thinking*, pp. 162–164, Harper & Brothers, New York, 1928.

[2] See Chap. X, Sec. 8.

and Very Able. Classification of instances into categories like these should be done independently by two or more persons as a check. If there is good agreement in the placing of individual instances, the percentages of the instances put in each given category by several judges may then be averaged to improve the accuracy. *Self-ratings* may be used, as well as ratings by others.

6. Scoring.—In the case of most score cards, the experimenter decides impressionistically what subscore, usually a percentage, should be given to each aspect of a variable (*e.g.*, the socio-economic status of a home). In other cases, a subscore is determined by counting the number of a certain item present in each instance (*e.g.*, books in a home), or by measurement in the stricter sense (*e.g.*, annual family income in dollars). The total score is the sum of the subscores on the different items included in the card. Usually the equality of the units, the placing of the zero point, the weightings, and the meaning of the total score are open to question; but in any case the total score represents a series of accumulated judgments reduced to a numerical common denominator. Scoring devices may be quite elaborate, as may be seen by inspecting Chapin's living room "scale" for scoring the socioeconomic status of a home, the Stanford-Binet intelligence test, or score cards for, say, dairy cattle used in judging contests at livestock shows. To show that they are parts of the same or associated things, the score on each item included on a score card should as a rule be high when the total score on the card is high, low when the latter is low. The theory of the score card is that the total score is an index or function of (varies with) the amount of the quality it is attempting to measure. Part of a living room score card designed by F. Stuart Chapin to measure the socioeconomic status of American homes is reproduced below.

CHAPIN'S SCALE FOR RATING LIVING ROOM EQUIPMENT[1]

DIRECTIONS TO VISITOR

1. The following list of items is for the guidance of the recorder. Not all of the features listed will be found in any one home. Entries on the schedules should, however, follow the order and numbering indicated.

---

[1] F. STUART CHAPIN, Scale for Rating Living Room Equipment, *American Journal of Sociology*, Vol. 37, pp. 583, 584, 1932.

Weights appear after the names of the respective items. Disregard these weights in recording. Only when the list is finally checked should the individual items be multiplied by these weights and the sum of the weighted score be computed, and then only after leaving the home. All information is confidential.

2. Check or underline the articles or items present. If more than one, write 2, 3, or 4, as the case may be.

3. Do not enter the *score* of any article or feature present. Complete recording before attempting to enter scores.

4. In cases where the family has no real living room, but uses the room at nights as a bedroom, or during the day as a kitchen or as a dining room, or as both, *in addition to use of room as the chief gathering place of the family, please note this fact clearly* and describe for what purposes the room is used.

5. When possible, it is desirable to have a living room checked twice. This may be done in either of two ways.

  a. After an interval of two or three weeks the same visitor may recheck the room. The first schedule should be marked I, the second II.

  b. After an interval or simultaneously, the room may be checked by two different visitors. One schedule should be marked *A*, the other *B*.

Scores of the same homes on two trials should be similar. If a group of homes are scored twice there should be a high correlation between the scores. Please report findings to F. Stuart Chapin, University of Minnesota.

<div align="center">Schedule of Living Room Equipment</div>

I. Fixed Features
    1. Floor.............. _____
       Softwood 1, hardwood 2, composition 3, stone 4.
    2. Floor covering..... _____
       Composition 1, carpet 2, small rugs 3, large rug 4, Oriental rug 6.
    3. Wall covering..... _____
       Paper 1, calcimine 2, plain paint 3, decorative paint 4, wooden panels 5.

    4. Woodwork........ _____
       Painted 1, varnished 2, stained 3, oiled 4.
    5. Door protection... _____
       Screen 1, storm door 1.
    6. Windows
       1 each window.... _____
    7. Window protection[1] _____
       Screen, blind, netting, storm sash, awning, shutter, 1 each.

[1] If checked out of season, ascertain if used in season and so record.

SCHEDULE OF LIVING ROOM EQUIPMENT.—*(Continued)*

8. Window covering[1]. _____
   Shades 1, curtains 2, drapes 3.
9. Fireplace......... _____
   Imitation 1, gas 2, wood 4, coal 4.
10. Fire utensils....... _____
    Andirons, screen, poker, tongs, shovel, brush, hod, basket, rack, 1 each.
11. Heat............. _____
    Stove 1, hot air 2, steam 3, hot water 4.
12. Artificial light..... _____
    Kerosene 1, gas 2, electric 3.
13. Artificial ventilators 1............ _____
14. Clothes closets 1.. _____
    Total section I.... _____

II. Built-in Features
15. Book containers... _____
    Shelves 1, cases 2.
16. Beds............ _____
    In a sideboard 1, in a ceiling 2, in a door 3.
17. Desk........... _____
18. Window seats 1... _____
19. Window boxes 1... _____
    Total section II... _____

III. Standard Furniture
20. Table............ _____
    Sewing 1, writing 1, card 1, library, end, tea, 2 each.
21. Chair............ _____
    Straight, rocker, arm-chair, high chair, 1 each.
22. Stool or bench.... _____
    High stool, footstool, piano stool, piano bench, 1 each.
23. Couch........... _____
    Cot 1, sanitary couch 2, chaise longue 3, daybed 4, davenport 5, bed-davenport 6.
24. Desk............. _____
    Business 1, personal-social 2.
25. Bookcases 1....... _____
26. Wardrobe or movable cabinet 1..... _____
27. Sewing cabinet 1.. _____
28. Sewing machine... _____
    Hand power 1, foot power 2, electric 3.

Etc., etc.

7. The Scale.—The ideal measuring device is the *scale*. By a scale is meant a sequence of interchangeable external units numbered from zero, such as a straightedge marked off into feet and inches. In sociology and psychology most attempts to develop scales have started from ranks or ratings. One of the simplest devices is the so-called *graphic rating scale*. The following is an example.

0	25	50	75	100
Completely submissive	Submissive	Average	Dominating	Completely dominating
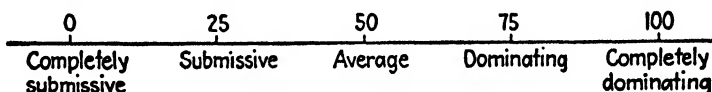
FIG. 1.—A simple graphic rating scale.

[1] If checked out of season, ascertain if used in season and so record.

Each judge rates each subject on a separate scale by making a mark on the scale where he thinks the subject falls. The distance of the mark from "Completely submissive" taken as zero is then measured in units of the spatial scale. The final rating of each subject is the average of the ratings given him by the several judges, provided there is a tendency toward agreement among them. The scale may become more objective, however, if a subject is scored, say, "80 per cent dominating" because he is observed to dominate (as tangibly defined) in 80 per cent of his contacts. This assumes that one contact is equal to another for the purpose in hand; but weighting may be applied if needed. Evidently, this kind of scale cannot claim the precision of scales in the physical sciences; but it is capable of very useful results.

If the ordinal[1] numbers derived from ranking are subjected to arithmetical treatment, such as addition or the calculation of means, it is implicitly assumed that the ranked instances are equally spaced on a linear scale. Thus, if we rank cities in respect to the efficiency of their governments, beginning with the least efficient, so that city C is 1, city A is 2, city B is 3, etc., and if we then use these ordinals as cardinals in arithmetical calculations, we imply that the government of city A is twice as efficient as that of city C, that the government of city B is 1.5 times as efficient as city A, etc. This assumption is, of course, inaccurate, but sometimes it is the best that can be done, or it is good enough for a particular problem. The zero point on such a scale is arbitrarily placed, usually coincident with or one unit below the lowest rank.

The most elaborate effort to build an exact scale yet made in the social sciences is probably that of L. L. Thurstone in the case of his scale for the measurement of an attitude, a sample of which is reproduced below.[2] Generalizing on Thurstone's method, and introducing minor modifications, it runs about as follows. A considerable number of supposed indexes of the attribute to be measured are chosen. Let us say that the attribute is "radicalism"; then the indexes might include membership in the Socialist party, admitted statements made against

---

[1] A cardinal number tells *how many* or *how much;* an ordinal number locates position in a series.

[2] L. L. THURSTONE and E. J. CHAVE, *The Measurement of Attitude,* University of Chicago Press, Chicago, 1929.

the existing social order, membership in a labor union, radical papers and journals read, expressed Communistic sympathies, signature on radical petitions, participation in strikes, jail sentences for radical activity, the authorship of radical articles and books, subscriptions to this or that radical doctrine, atheism, unconventional sexual behavior, and so on.   After these indexes have been selected and defined as objectively as possible, they are submitted to a number of qualified judges, who are asked to rank them in the order of the degree of radicalism that each seems to imply.   Indexes that appear to indicate about the same degree of radicalism are regarded as ties.   The indexes are thus collected into successive piles, which to the judges should seem to be equally spaced apart in degree of radicalism.   When the judges have finished ranking the indexes, each index is assigned the *average* rank given it by the several judges, except that any index about which the judges differ too much is rejected entirely.

Index C   Index A   Index B
0 3 5   1011 15   20   80   85   90 92 95   100
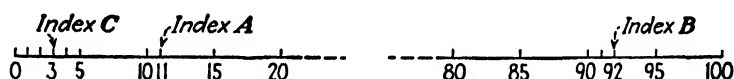
FIG. 2.—Diagram of a generalized Thurstone attitude scale.

Each index will then have an average rank or scale value, and these values may if desired be converted to a percentage scale, from the lowest value taken as zero to the highest value taken as 100 (see Fig. 2).   The scale is then ready to be applied to other samples of instances (say persons), by simply checking on a list of the indexes those that apply to a given individual, adding the scale values of the indexes checked, and averaging them.   Each individual may thus be given a scale value that is supposed to measure in a relative way the amount, say, of "radicalism" that characterizes him.

Thurstone's attitude scale has often given results that correlated highly with those obtained by much simpler procedures, such as graphic rating scales, and ratings[1] or rankings represented by consecutive numbers.   It has also been criticized on various theoretical grounds.[2]

[1] For example, individuals are classified as Very Radical, Radical, Neutral, Conservative, Very Conservative, and those in the Very Radical group are given a score of one, those in the Radical group a score of two, etc.

[2] See R. K. MERTON, Fact and Factitiousness in Ethnic Opinionnaires, *American Sociological Review*, Vol. 5, pp. 13–28, 1940.

SAMPLE OF A THURSTONE ATTITUDE SCALE[1]

EXPERIMENTAL STUDY OF ATTITUDE TOWARD THE CHURCH

· ·

Check (✔) every statement below that expresses your sentiment toward the church.   Interpret the statements in accordance with your own experience with churches.

	Scale value
1. I think the teaching of the church is altogether too superficial to have much social significance.........................	8.3
2. I feel the church services give me inspiration and help me to live up to my best during the following week.............	1.7
3. I think the church keeps business and politics up to a higher standard than they would otherwise tend to maintain......	2.6
4. I find the services of the church both restful and inspiring...	2.3
5. When I go to church I enjoy a fine ritual service and good music..	4.0
6. I believe in what the church teaches but with material reservation...	4.5
7. I do not receive any benefit from attending church services but I think it helps some people........................	5.7
8. I believe in religion but I seldom go to church............	5.4
9. I am careless about religion and church relationships but I would not like to see my attitude become general..........	4.7
10. I regard the church as a static, crystallized institution and as such it is unwholesome and detrimental to society and the individual..	10.5
11. I believe church membership is almost essential to living life at its best..	1.5
12. I do not understand the dogmas or creeds of the church but I find that the church helps me to be more honest and creditable...	3.1
13. The paternal and benevolent attitude of the church is quite distasteful to me....................................	8.2
14. I feel that church attendance is a fair index of the nation's morality...	2.6
15. Sometimes I feel that the church and religion are necessary and sometimes I doubt it..............................	5.6
16. I believe the church is fundamentally sound but some of its adherents have given it a bad name....................	3.9
17. I think the church is a parasite on society...............	11.00

There are also several important methods of converting ranks to a scale having more or less equal units and an arbitrary zero point that are outside the scope of this text.[1]   Probably the most scientific is the mathematical method of curve fitting.[2]

When the concepts of space, time, money, weight, mass, and so on, are used in sociology, they are of course amenable to accurate measurement by scales already scientifically established.

8. Discrete Aggregates.—Population aggregates are of great importance in sociological studies.   It is possible to define these aggregates (communities, neighborhoods, families, and the like) so that their number can be counted.   We hold that it is also possible to measure the size of such aggregates by counting the number of individuals that compose them.   We do this in the belief that the only essentials of measurement are units that are equal and interchangeable *for a purpose*.   The sociologist finds it more useful for his purposes to measure the size of the family in terms of the number of its members than in terms of their weight in pounds or their height in inches.   The nature of a "member" does not vary from person to person in any way that interferes with the purpose.   Moreover, since there is no point in subdividing a "member," nothing is lost because it is logically a discrete unit.   This idea of measurement can also be extended to any other sociological concept that can be broken down into parts that are equal and interchangeable for the purpose in hand.

9. The Measurement of an Intangible Quality.—All attempts to measure an intangible quality, such as an attitude, must, of course, be indirect in type.   The classic example of indirect measurement in the physical sciences is a thermometer that uses the changing length of a column of mercury as an index of change in the amount of the intangible quality "temperature."   In the case of the indirect measurement of a quality Y (temperature) in terms of an index X (mercury column), there should ideally

---

[1] See J. P. GUILFORD, *Psychometric Methods*, McGraw-Hill Book Company, Inc., New York, 1936; P. M. SYMONDS, *Diagnosing Personality and Conduct*, pp. 86–89, D. Appleton-Century Company, Inc., New York, 1931.

[2] KARL J. HOLZINGER, *Statistical Methods for Students in Education*, pp. 221–224, Ginn and Company, Boston, 1928; C. H. RICHARDSON, *An Introduction to Statistical Analysis*, Chaps. VIII and X, Harcourt, Brace and Company, Inc., New York, 1934.

be a perfect straight-line relationship between the two (see Chap. X), so that each unit change in X represents a constant amount of change in Y. But since Y is an intangible and cannot be directly measured, there is no way of proving that such a relationship exists between Y and X. So, while we may be certain that a scale distance of say $4X$ is twice as great as a scale distance of $2X$, we cannot be certain that a scale distance of $4X$ represents twice as much of Y as does a scale distance of $2X$. A child with an I.Q. of 120 is probably not just twice as intelligent as another child with an I.Q. of 60. All devices of indirect measurement, including the thermometer, are open to this objection. But the scientific and practical usefulness of the thermometer and of other indirect measuring devices suggests that for many purposes this is not serious. Usually the important things are rather that the same absolute reading on the X scale shall always represent the same amount of the intangible quality Y, as verified by introspection or by some external result in which we are interested (*e.g.*, at 32°F., water freezes); and that the X scale shall be able to differentiate changes in Y small enough for our purposes. We shall then know what to expect from Y when the scale registers a certain value of X. If the relationship is close enough to permit a useful prediction of Y from the reading on the X scale, the latter may still be valuable and is not to be discarded until a better index is found.

In practical scale or score-card making, where there is an attempt to measure an intangible quality Y in terms of a tangible index X, it is often helpful to set up a "fundamental interval" for subdivision. This is done by selecting two extreme observable instances of Y, marking the values of X corresponding to them "0" and, say, "100" respectively, and dividing the included range of X into 100 equal units. In the case of one thermometer, the extreme instances of temperature are taken at the melting point of ice and at the condensing point of steam. As a parallel, in mental testing, for certain purposes we might regard inability to pass the first grade in school as indicative of zero intelligence, and ability to finish the university with honors as indicative of 100 per cent intelligence, and represent intermediate degrees of intelligence by scores between 0 and 100. As with most thermometers, for many purposes the zero point need not denote an absolute zero, and the upper limit need not mean the ultimate

maximum amount of Y. It is important, however, to make sure that the "fundamental interval" includes as large a range of data as investigators will require.

When the quality Y is subjective (*e.g.*, happiness in marriage), it has already been implied that there are two ways of testing the amount of relationship between it and the tangible index X (*e.g.*, a score on the Burgess-Cottrell[1] scale for measuring happiness in marriage): (1) by comparing the amount of Y indicated by the X instrument with the subjective judgment of the subject or of a competent observer (*e.g.*, couples getting high scores on the Burgess-Cottrell scale *consider* themselves happy)—this is appropriate if interest centers in the subjective quality as such; and (2) by checking the readings of the X instrument against certain tangible conditions that are ascribed to Y (*e.g.*, low happiness scores on the Burgess-Cottrell scale are followed by divorce more often than are high scores). These are called tests of validity. Validity is also established in part by definition and agreement, *e.g.*, the cooperative definition described in Chap. IV. Chapin's living room scale, mentioned above, is intended to measure the socioeconomic status of the homes to which it is applied. The fact that the card has given higher scores when applied to upper middle class homes than when applied to middle class homes, determined independently, is evidence of its validity. Its reliability was established when different observers used it on the same homes with little variation in results.

Evidently, the indirect measurement of a subjective quality must wait upon the discovery of a satisfactory tangible index, which is to be sought among the apparent results or causes of the subjective quality, among the results of common causes, or, from a different point of view, among the external aspects of the subjective concept. Thus, the expansion and contraction of the column of mercury in a thermometer are apparently the *result* of changes in temperature.

Whether the measurement of an intangible quality by means of a tangible index or by means of introspective ratings converted to scale values is superior depends upon particular circumstances, and especially upon the direction of interest. If possible, both should be carried through for purposes of validation.

[1] E. W. BURGESS and LEONARD J. COTTRELL, *Predicting Success or Failure in Marriage*, Prentice-Hall Inc., New York, 1939.

10. Rules of Measurement.—We summarize below what are probably the most useful rules of measurement in social research.

1. The quality that it is desired to measure should be defined verbally as clearly as possible in the beginning. But the measurement of a quality is also a crucial part of its definition. In fact, "what the scale measures" may later be regarded as preferable to the verbal definition, as equivalent to it, or as not at all equivalent to it, depending on the degree of validity established for the scale and the usefulness of its results.

2. The purpose of the measurement should be stated or understood.

3. The unit used should be appropriate to the purpose of the measurement.

4. Units should be equivalent one to another (equal, interchangeable) *for the purpose in view;* except that in the indirect measurement of an intangible quality in terms of a tangible index the equality of the intangible units is indeterminate, and for many purposes is unimportant.

If the units of a scale are sufficiently equal for a purpose, it is safe for that purpose to add or average them, to interchange them, or to claim that, say, two units represent twice as much of the quality as does one unit.

For the historian, one year is not equivalent to another; for the actuary constructing a life table, it is.

5. The unit should be applied as exclusively as possible to the quality defined for measurement, in accordance with the purpose stated.

That is, in measuring a man's height in inches, we should not include his shoes, nor should we measure him in a slouched posture. So, in measuring "intelligence," we should, if possible, exclude inequalities of effort.

6. The unit should be applied to the entire range in which the investigator is interested.

In applying an inch end-over-end, or an inch scale, to measure the height of a man, no part of the total distance that is his height should be skipped or measured in other than a single straight line. When a Fahrenheit thermometer registers the temperature, however, it reads above or below a fixed point that is arbitrarily called zero. This is adequate for ordinary

purposes, because most of us are interested only in the range of temperature included in the thermometer, and not in an extension of that range to a depth never observed in ordinary experience. But for some scientific work, the temperature needs to be measured from a true zero point, and a different scale is used.

The ratio of two measurements holds only with reference to the zero point from which they are made. If this is not an absolute zero, that fact should not be forgotten when interpreting the ratios.

7. The size of the unit should be fine enough to detect the smallest differences that are of importance for the inquiry, but need be no finer.

8. Final judgment of an instrument designed to measure an intangible quality should depend chiefly on tests of its *validity* and *reliability*.

Summary.—We have seen that even "subjective" qualities are amenable to a great deal of statistical analysis through counting, classification, ranking, and rating. They cannot be exactly measured unless the form of their theoretical distribution is known a priori, or unless they are perfectly correlated with some objective index; and it is seldom or never possible to demonstrate completely either of these propositions. Nevertheless, such qualities have already been measured in both the natural and the social sciences successfully enough to satisfy many important scientific and practical uses. Devices like the Binet test and like those used to score social attitudes, socioeconomic status, personality traits, and so on, are promising approaches to measurement in social research, and their rapid improvement and extension to cover many more sociological concepts are to be anticipated. Moreover, objective qualities in which sociology is interested not only can be counted, classified, and the like, but they can also either be measured by the scales already standardized by the physical sciences or they should offer no difficulties that are peculiar to the social sciences.

Exercises

1. Is anything more than clearness of definition necessary to render data amenable to statistical treatment? Illustrate.

2. Can classification and counting alone form any basis for statistical analysis? Illustrate.

3. What are the main points to watch in the use of classification? Illustrate.

4. How does classification differ from rating?   Illustrate.

5. Give an example of the kind and amount of ability a judge should have to qualify as a "rater."

6. Name at least one method of converting ranks to scale values.

7. Devise a simple graphic rating scale for the personality trait of "sociability."

8. Describe some scoring device used in sociology.   What is your opinion of it as a measuring instrument?

9. Distinguish between a scoring device and a scale in the strict mathematical sense.

10. Distinguish between counting and measurement.

11. Illustrate a sociological problem where counting is equivalent to measurement.

12. Discuss the possibility and necessity of equal units in the measurement of an intangible quality.

13. What is of chief importance in the indirect measurement of an intangible quality?

14. What is meant by the validity of a measuring scale?   By its reliability?   How can an instrument designed to measure an intangible quality be validated?   Illustrate.

15. Give an example of an intangible quality of interest to sociology, and describe briefly two ways in which it may be measured.

16. What method of measurement would you apply to answer each of the following questions:

   a. Does *divorce* tend to increase with family income?

   b. Do the *ablest* people leave the farm for the city?

   c. How do 10 cities compare in respect to *good government?*

17. What is the reason for taking a number of measurements of the same thing and averaging them?

References

CAMPBELL, N. R.: *An Account of the Principles of Measurement and Calculation*, Longmans, Green & Company, New York, 1928.

CHAPIN, F. STUART: Measurement in Sociology, *The American Journal of Sociology*, Vol. 40, pp. 426–480, 1935.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chaps. I and VII, Prentice-Hall, Inc., New York, 1939.

JOHNSON, H. M.: Pseudo-mathematics in the Mental and Social Sciences, *American Journal of Psychology*, Vol. 48, pp. 342 ff., 1936.

KIRKPATRICK, CLIFFORD: Assumptions and Methods in Attitude Measurement, *American Sociological Review*, Vol. 1, pp. 75 ff., 1936.

LUNDBERG, G. A.: The Measurement of Socio-economic Status, *American Sociological Review*, Vol. 5, pp. 29 ff., 1940.

SCATES, DOUGLAS: The Essential Conditions of Measurement, *Psychometrika*, Vol. 2, pp. 27 ff., 1937.

TERMAN, LEWIS M., and MAUD A. MERRILL: *Measuring Intelligence*, Houghton Mifflin Company, Boston, 1937.

CHAPTER III

FACTOR CONTROL

Among the social sciences the controlled experiment has been employed much less than in the natural sciences. As a rule, sociologists have either preferred or felt obliged to investigate social situations in all their original complexity and confusion. The methods for dealing with this kind of data attempt to introduce control by means of classification in the case of *attributes* (unmeasured traits, *e.g.*, married, single) and by mathematical devices in the case of *variables*, (measured traits, *e.g.*, age in years).

1. The Actuarial Method.—One of the most effective schemes of classifying attributes is similar in general principle to that employed by actuaries in determining insurance risks.[1] For example, a large number of paroled criminals may be sorted into relatively homogeneous groups with respect to various criteria, such as number of previous arrests, prison record, age, type of offense committed, intelligence, and so on, and the rate of violation of parole determined for each group. After proper testing, these rates may then be used as estimates of the probability of violation of other prisoners who fall in the established classifications.

We begin with a specified group of items, say paroled prisoners from the Joliet (Ill.) penitentiary on Jan. 1, 1941. The simplest classification is a *dichotomy*, or separation of the A's from the *Not A*'s. Thus, our parolees may be divided into the married and the not married. If we wish to test whether marital status (trait A) is associated with success on parole (trait B), we compare the proportion of successful parolees (B's) among the married parolees (A's) with the proportion among the not married parolees (*not A*'s). When there is no association, *i.e.*, the traits A and B are independent, the two

[1] For a more thorough development of this technique, see G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Chaps. I–V, Charles Griffin & Company, Ltd., London, 1937.

proportions will be the same, except for chance errors. In other words, if 80 per cent of the married parolees succeeded, but only 60 per cent of the not married parolees did so, we would conclude that marital status was favorable to success on parole.

Suppose we believe that a good prison record (trait C) also makes for success on parole. We test it in the same way as we did marital status above, and confirm our belief. It may then be worth while to make a double classification of the parolees by marital status and by prison record, as shown in Table 1. From this table we note that the proportion of successful parolees in the group as a whole is $\frac{360}{500} = 0.72$, among the married is $\frac{240}{300} = 0.80$, and among the married with a good prison record is $\frac{65}{70} = 0.93$, approximately. On the other hand, among the

TABLE 1.—CLASSIFICATION OF 500 PAROLEES BY MARITAL STATUS AND PRISON RECORD, JOLIET, ILL., JAN. 1, 1941. (HYPOTHETICAL DATA)

Outcome	Parolees, married		Parolees, not married		Total
	Record good	Record not good	Record good	Record not good	
Successful.............	65	175	25	95	360
Not successful..........	5	55	10	70	140
Total...............	70	230	35	165	500

not married parolees with a not good prison record, the proportion of successes is $\frac{95}{165} = 0.58$ nearly. Evidently, in future groups of parolees chosen in the same way and exposed to the same general conditions as were the 500 represented in Table 1, a married man with a good prison record may be expected to have a much better chance of succeeding than a man not married with a prison record that is not good. More specifically, for every man of the first type that failed, we should expect 6 of the second type to fail, out of equal numbers placed on parole.

It is, of course, possible to subclassify the cases in Table 1 still further, either by substituting more complete breakdowns for the dichotomies (*e.g.*, married, single, divorced, widowed for married, not married), or by introducing additional factors (*e.g.*, employment record before arrest).[1]

[1] See Chap. XI.

2. The Search for Causes.—It is often said that the underlying purpose of all science is prediction.   Certainly, scientific research constantly seeks to discover causes.   Much philosophical dispute has occurred regarding the nature and reality of a cause, but we shall here say only that we mean by a cause any factor whose change under controlled conditions is invariably followed or accompanied by a change in a second factor.   The logicians refer to this as *concomitant variation*.   The kind of causes with which practical science is most concerned are simply factors that give the easiest and most reliable prediction, or understanding, of certain conditions that constitute a problem. Thus, if we can always change the divorce rate in a given type of social situation by changing the proportion of Protestant-Catholic marriages, the intermarriage of Protestants and Catholics may be regarded as one cause of divorce under the given conditions.   In the social sciences, there are always many causes that combine to produce any actual situation or result. Evidently the divorce rate of a city is the product of a vast number of forces, only some of which can be discovered or controlled.

3. Matching Experimental and Control Groups.—The logical requirements for establishing a causal relationship are the same in every science.[1]   It is always necessary to establish the fact of concomitant variation.   For working purposes, the procedure is essentially to introduce, remove, or vary in amount the suspected cause, and then to observe or measure the corresponding changes, if any, in the thing that is expected to be affected. For example, suppose that we want to test the belief that knowledge of the evils of alcohol will prevent young people from drinking.   We expose a number of such persons to appropriate instruction and note what proportion of them acquire the habit of drinking within, say, a two-year period.   In this group, called the *experimental group*, the supposed cause is present.   A second group of young people, which may be termed the *control group*, is given no instruction, so that the supposed cause is absent. After two years, the proportion of habitual drinkers is determined in this group also, and the proportions are compared between the experimental and control groups.   If the experi-

[1] See John Dewey, *Logic: The Theory of Inquiry*, pp. 101, 462, 491, 509, Henry Holt and Company, Inc., New York, 1938.

mental group shows a lower percentage of drinkers than the control group, however, it still cannot be said that the instruction made the difference, unless it can also be shown that nothing else is likely to have done so. Thus, it is possible that the experimental group contained a considerably larger proportion of women or of church members than the control group, which might make the comparison unfair. It is evidently necessary in any experiment that the experimental and control groups shall be essentially alike in all important respects that might affect the outcome, except for the factor or factors under investigation. This must, of course, be taken care of when the experiment or investigation is being planned. The young people in our experimental group must have no characteristics, except the instruction, that will make them more liable or less liable to become drinkers than those in the control group. The usual way of trying to insure this equality is to *match* the two groups in respect to every important point that may be related to drinking, such as age, sex, family background, church membership, present drinking habits and attitudes, and so on. Moreover, all conditions must remain approximately the same for the two groups during the two years that the experiment is under way.

4. The Principle of Randomization.—In sociological research, however, it is seldom that an investigator can feel that his experimental and control groups are actually matched in all important respects needed to insure a valid comparison between them. He is, therefore, obliged to summon to his aid the principle of *randomization*. Having matched his two groups as well as he reasonably can, he then decides by a random draw which of each pair of matched subjects, or which subjects from the total lot, shall belong to the experimental group and which to the control group. If this is not feasible, it may be decided by a draw which of the two matched groups shall be the experimental one, or this may be done in addition to the above. As long as there are only two groups, this latter method of randomization alone is not very effective. The experiment will be better designed if there can be several groups, or *replications*, half of which are drawn at random to serve as the experimental groups. In some cases, indeed, the whole process of matching the groups may best be omitted, and dependence placed in subdividing the potential events—*e.g.*, a large number of unselected young

people—into two or more groups by random selection. When any good method of randomization is used, all initial differences between the experimental and control groups should be accidents of chance.[1]

5. Pretests and Final Tests.—Whatever method of equalization is used, it is well before subjecting the groups to the conditions of the experiment to test them to see how much alike the experimental and control groups really are in pertinent respects. This is usually done by means of a pretest, which is the same as the final test that will be used at the end of the experiment to measure the differences between the groups at that time. Thus, in our illustration, we might set up a battery of questions about drinking habits that would enable us to decide to what extent a young person drank or was predisposed to drink, and if the experimental and control groups scored about the same on this test, we might regard them as equivalent for the purposes of our investigation.

6. The Influence of Additional Factors.—It is often desirable to test the effects of a third factor on the relationship between the independent and dependent factors in an experiment. In this case, the third factor is inserted and removed, with only the independent and dependent factors present. Thus, we might observe the influence of sex in studying the influence of instruction on drinking. Both control and experimental groups would then be divided by sex, giving four groups rather than two.

7. The Case of Continuous Variables.—In the illustration above, we were dealing with attributes, such as "instruction," "no instruction," "habitual drinkers," "not habitual drinkers," rather than with measured variables, like the amount of instruction and the amount of the tendency to drink. Although there is no difference in principle between the two cases, there is some variation in procedure. Thus, if we wanted to measure the amount of the tendency to drink in relation to the amount of instruction given, we should take several groups instead of only two. To each of the several groups we should give a different *amount* of instruction, including no instruction at all

[1] For a more advanced discussion of this subject, together with the statistical techniques of analysis of variance and covariance that have recently been developed in connection with it, see E. F. Lindquist, *Statistical Analysis in Educational Research*, Chaps. IV–VI, Houghton Mifflin Company, Boston, 1940.

to one group, and note whether there was any relationship between the increasing amount of instruction and the tendency to drink after two years. As before, we should have to equate the groups in all important respects before experimenting with them, or else be prepared to make corrections for the differences. Of course, we should have to devise scales for measuring the amount of instruction and the amount of the tendency to drink, before we could treat these factors as continuous variables.

8. Interfering Variables.—As in the case of attributes above, it is usually important to measure the influence of certain interfering variables. In our drinking experiment, some of these might be the attitude of the parents toward drinking, the subjects' ages, their money incomes, and so on. Such variables are not matched or randomized out of the experiment, but are introduced in varying known amounts, and their effects on the independent and dependent variables are measured. Factors may then be held constant, or their influence subtracted out, by mathematical methods.[1] This type of analysis yields more information and information of a more practical kind than when all interfering factors are actually removed by matching or are equalized by randomization; and it is also generally easier to carry out.

Exercises

1. Illustrate the use of the actuarial technique in the prediction of success in marriage.

2. Explain how you would obtain control over interfering factors in a study designed to show the effects of the presence of children on the divorce rate, or other problem of your choosing.

3. Comment briefly on the following published statements:

a. "Despite marked advances in appendicitis diagnosis and surgery, Wisconsin's death rate from the ailment, which stood at 11.6 deaths per 1,000 population in 1911, nevertheless increased to a rate of 18.2 in 1930."[2]

b. "*Women Are Safer Drivers than Men Records Reveal:* When Mary and Jack borrow Dad's car for a ride, they'll be smart if they let Mary do the driving.

---

[1] See, for example, Mordecai Ezekiel, *Methods of Correlation Analysis*, Chap. XIII, John Wiley & Sons, Inc., New York, 1930; or G. W. Snedecor, *Statistical Methods*, rev. ed., Chaps. XII and XIII, Collegiate Press, Inc., of Iowa State College, Ames, Iowa, 1938.

[2] *Wisconsin State Board of Health Bulletin*, Madison, April–June, 1935, p. 26.

"For in spite of the young man's claim to being a better driver, state highway commission records show that women drivers seldom are involved in fatal accidents. Young men, however, are involved in more fatal automobile crashes than any other age class of motorists.

"Few women drivers are found on state highway commission fatality records, and only one person was killed in the last two years by a girl driver under 18 years of age.

"State safety workers won't argue that Mary is a better driver than Jack, but they do claim that state records indicate she is a safer driver."

c. "*Homemaking Careers Attracting More Girls:* In increasing number, girls are turning attention these days to homemaking as a career.

"The popularity of homemaking courses is shown in the increasing enrollment in home economics at the University of Wisconsin where enrollment this fall is nearly 10 per cent above 1936, according to the director of the course."

d. "There has been more social progress in the United States in the last 18 years since women have had the vote."

e. "The Distilled Spirits Institute, demanding that the Anti-Saloon League recognize the prevailing downward trend of major crimes, bases its case largely on this general statement: The total (of all crimes) for the calendar year 1936 showed a decrease of 112,055 offenses as compared with 1935."

(Turn in to the instructor two examples of the misuse of statistical reasoning clipped from newspaper or magazine.)

References

BURTT, E. A.: *Principles and Problems of Right Thinking*, Part III, Harper & Brothers, New York, 1928.

CHADDOCK, R. E.: *Principles and Methods of Statistics*, Chaps. II and III, Houghton Mifflin Company, Boston, 1925.

CHAPIN, F. STUART: An Experiment on the Social Effects of Good Housing, *American Sociological Review*, Vol. 5, pp. 868–879, 1940.

DEWEY, JOHN: *Logic: The Theory of Inquiry*, especially Chaps. XI and XXIV, Henry Holt and Company, Inc., New York, 1938.

FISHER, R. A.: *The Design of Experiments*, D. Van Nostrand Company, Inc., New York, 1935.

GOOD, C. V., A. S. BARR, and D. E. SCATES: *The Methodology of Educational Research*, Chaps. IX and X, D. Appleton-Century Company, Inc., New York, 1936.

GOULDEN, C. H.: *Methods of Statistical Analysis*, Chaps. I and V, John Wiley & Sons, Inc., New York, 1939.

PETERS, C. C., and W. R. VAN VOORHIS: *Statistical Procedures and Their Mathematical Bases*, Chap. XVI, McGraw-Hill Book Company, Inc., New York, 1940.

WOLF, A.: *Essentials of Scientific Method*, The Macmillan Company, New York (no date).

CHAPTER IV

THE STATISTICAL INQUIRY

1. The Role of Nonquantitative Methods.—Access to nonquantitative methods, such as the historical method, the case study, and the general interview, is not to be denied the statistical investigator in sociology. Many of his problems and ideas will be suggested by working with materials of these kinds before the statistical study is set up. Also, during the progress of the collection of the statistical data and analysis of them, he will usually find it invaluable to interview or talk with the informants and their neighbors, to saturate himself with their points of view and backgrounds, and to judge the reliability of their replies to formal schedule questions by shrewd observation. Finally, as suggested in Chap. I, in interpreting his statistical findings, some important questions are almost certain to arise that cannot be answered from the figures in hand, and he will want to go back to the living situations for fresh suggestions. The statistical investigator is expected, however, to limit his formal conclusions to those arrived at by tested quantitative methods.

2. The Problem.—The statistical problem in sociological research may vary from what is exploratory and merely fact finding to the testing of a sharply stated hypothesis, depending upon how much is already known about the subject. We may set up a study to find out anything we can about divorce in the United States, or we may limit the inquiry to testing the hypothesis that the occupation of the husband plays an important part in the situation. Exploratory or fact-finding studies should be regarded as merely preliminary to more specific and better controlled studies, because the former cannot penetrate beneath the surface of social phenomena. The problem should also be cut to fit the limitations of time, money, and personnel qualifications at the disposal of the investigator. It should usually be a problem of obvious theoretical or practical impor-

tance, although a certain amount of research without apparent value but of interest to the investigator should be encouraged, because this kind of probing about has sometimes resulted in important scientific discoveries. The availability or lack of availability of reliable statistical data is another consideration that will affect the choice of a problem. This bears on the point that the problem must be capable of quantification or measurement. Above all, the problem should lie in the field of methodological and informational competence of the investigator, but as far as possible outside his field of personal bias. There is sometimes a conflict here, as when a Negro sociologist wishes to investigate the social conditions of the Negro race. He should know the field better for being a Negro, but he is likely to carry into the study a racial sympathy that may influence his findings. It is very desirable for an investigator to state frankly his biases, as well as to do his best to overcome them.

Of course, no problem should be finally selected until it is known to what extent and by what methods it has already been studied.[1] Although some investigations need to be repeated or done differently for confirmation, it sometimes happens that a problem has been very satisfactorily solved, and further work on it would be a waste of time. What is more likely is that certain angles of the problem have been worked out, but other angles remain to be investigated. The research worker is, therefore, guided by a knowledge of previous work into the most profitable channels for further study, and may obtain suggestions and warnings from what others have done.

In dealing with a statistical problem of the more scientific sort, it is indispensable to state the problem as a formal hypothesis or hypotheses to be tested. Such a hypothesis should be so worded that the task of the investigator is made as easy as

---

[1] Aids in locating previous sociological research on a topic include the files of *The American Journal of Sociology, The American Sociological Review, The Journal of Social Forces, Sociology and Social Research,* and *Population Index; Social Science Abstracts* (1929–1932); P. K. Whelpton, *Needed Population Research,* Science Press Printing Company, Lancaster, Pennsylvania, 1938; *The Psychological Index; Encyclopedia of the Social Sciences,* E. R. A. Seligman, ed., The Macmillan Company, New York, 1930; *Poole's Index to Periodical Literature; Readers' Guide to Periodic Literature; Annual Magazine Subject Index; Book Review Digest; United States Catalog: Books in Print; Cumulative Book Index.*

possible. It is usually simpler to use a positive hypothesis than a negative one, and then to try to disprove rather than to prove it. Strictly speaking, we can never prove a general affirmative proposition because we cannot examine all possible cases; but a single exception may effectively disprove it. Thus we might take as a hypothesis, "Any association found between the birth rate and the business index, with the marriage rate held constant, is due to chance errors," and seek to show that in our particular sample it is *not* due to chance errors. We can only disprove, or fail to disprove, such a hypothesis. For practical purposes, however, we may regard as provisionally true any hypothesis that careful tests have failed to disprove.

3. Secondary Statistical Data.—Research is a cooperative social enterprise, and the social investigator often necessarily uses data collected by someone else. The chief sources of secondary statistical data that are of interest to sociologists are the publications of the various bureaus and divisions of the Federal, state, county, and municipal governments, and a few private agencies.[1]

[1] Important Federal agencies in the United States include the Bureau of the Census, the Division of Rural Life and Welfare of the Department of Agriculture, the Bureau of Agricultural Economics, the Bureau of Labor Statistics, the Children's Bureau, Public Health Service, Works Projects Administration, Division of Vital Statistics of the Bureau of the Census, National Resources Committee, Interstate Commerce Commission, Central Statistical Board, Department of Commerce, Department of the Interior, Federal Bureau of Investigation, National Archives, National Youth Administration, Tennessee Valley Authority, Women's Bureau, United States Employment Service, Immigration and Naturalization Service, Agricultural Adjustment Administration, Farm Security Administration, Office of Education in the Department of the Interior, Office of Indian Affairs in the Department of the Interior. A current summary of Federal agencies, their subdivisions and activities, is available in the *United States Government Manual* issued by the National Emergency Council. A general source for the purchase of Federal documents is the Superintendent of Documents. All these agencies are located in Washington, D. C.

Information about births, marriages, divorces, deaths, and the public health is published by state bureaus of public health or vital statistics, with offices in the state capitals. State bureaus of correction, departments of education, departments of agriculture, departments of public welfare, planning boards, tax commissions, and the like are important sources of data for students of social conditions. State and private universities and agricultural colleges also gather and interpret a great deal of information. The

Any serious statistical research project will, of course, soon lead far beyond any general summary of sources of data. Much of the success of the trained investigator depends upon his ingenuity and persistence in discovering the available data that are pertinent to his problem. Intimate familiarity with the field of investigation is the best aid here.

After secondary data are found, however, the investigator must examine them carefully and critically before he can safely use them for his special purpose. He needs to know (1) the definition of the thing that is enumerated in relation to his purpose, or (2) the definition of the whole that is measured and of the unit by which it is measured, (3) the exhaustiveness and mutual exclusiveness of the classification, (4) changes in the definition, (5) the extent of actual over- or underenumeration or measurement, (6) the date or period in time to which the data apply.

A few examples may be of help. In the 1935 Census of Agriculture in the United States, a farm was carefully defined as

. . . all the land which is directly farmed by one person, either by his own labor alone or with the assistance of members of his household, or hired employees. A ranch, nursery, greenhouse, hatchery, feed lot, or apiary is considered a farm. Establishments keeping furbearing animals or game, fish hatcheries, stockyards, parks, etc., are not considered as farms unless combined with farm operations.

The enumerator was instructed *not* to report as a farm any tract of land of less than 3 acres, unless its agricultural products in 1934 were valued at $250 or more.

Brookings Institution of Washington, D. C., the National Bureau of Economic Research of New York, the Russell Sage Foundation of New York, the Scripps Foundation for Population Research of Oxford, Ohio, and the Gini Foundation of Palo Alto, Calif., are private organizations whose work is of value to social investigators.

The latest copies of the *Statistical Abstract of the United States*, published by the United States Department of Commerce; the Abstract of the Census of the United States, published by the United States Bureau of the Census; and the *World Almanac*, obtainable at most newsstands, are of frequent use. Bibliographies include those of Dorothy C. Culver, *Methodology of Social Research: A Bibliography*, and of A. F. Kuhlman, *Public Documents*.

The League of Nations, the International Labor Office, and the International Institute of Agriculture publish much statistical material of world interest, available in public libraries.

A farm may consist of a single tract of land, or of a number of separate tracts. These several tracts may be held under different tenures, as when one tract is owned by the farmer and another tract is rented by him. When a landowner has one or more tenants, croppers, or managers, the land operated by each is considered a farm. Thus on a plantation the land operated by each "cropper" or tenant was reported as a separate farm. The land operated by the owner or manager, by means of wage hands, was likewise reported as a separate farm.

That this definition of a "farm" nevertheless did not suit the purposes of all users of the census appears from comments like the following:

The census uses a concept of a "farm" which is an arbitrary statistical definition violating any sound reasoning from whatever standpoint we may choose. In counting farm operators the census makes no distinction between the sharecropper on the one hand, and, on the other hand, the farmer who operates his property either personally or with the aid of a manager and the tenant who operates a farm—strange as it may seem, in current American agricultural statistics the plantation does not exist. Paradoxically enough, it lives statistically under the disguise of its direct competitor and adversary, the small family farm . . . nobody knows how many plantations existed in the United States in 1920, 1925, 1930, or 1935.[1]

A great many more farms were enumerated by the Census of Agriculture in 1935 than in 1930. Between these two censuses no change was made in the definition of a farm; yet there is evidence that the 1935 census counted as farms many plots that were not counted as farms in 1930, especially in or near mining and industrial areas. The depression and unemployment caused the occupants of these plots to give more than ordinary attention to gardening, chicken raising, and other home production, and as a result these rural home places were lifted into the farm class. Since the families and the plots were otherwise just the same as they had been in 1930, and the "farmers" added by the 1935 census were actually miners and industrial workers who would return to their usual employment at the first opportunity, it has been felt that the heavy increase in the number of farms reported was largely spurious. As usual, however, the error, if it may be so called, occurred on the periphery of the definition

[1] KARL BRANDT, Fallacious Census Terminology and Its Consequences in Agriculture, *Social Research*, Vol. 5, pp. 19–37, 1938.

where the concept defined (a farm) shades off into something different (not a farm). Most of the farms added in the above manner were quite small, and the value of their products was so close to the minimum of $250 that they might easily slip in and out of the farm category. The number of farmers returned by the census of agriculture is never the same as the number found by the accompanying census of occupations.

In the case of farm laborers, including members of the farmer's family working on the home farm, the problem of definition is so difficult that not much reliance can be placed in the figures furnished by the census. In addition, the census of 1920 was taken as of Jan. 1 and that of 1930 as of Apr. 1, and this shift of date alone caused a sharp variation in the number of farm laborers reported. It is well known that the census of population underenumerates young children, Negroes, and other classes that for one reason or another are likely to be overlooked; that the reporting of the population by years of age overloads the 5's and 10's (*e.g.*, 15, 20), at the expense of the other years (*e. g.*, 14, 17, 19, 22); and so on.

Such examples suggest only a few of the many pitfalls that lie in secondary data, even when collected by a great national agency like the Bureau of the Census, which may be regarded as unbiased and thoroughly honest in those aspects of its work that cannot be checked by the consumer of the data. The dangers are usually much greater in the case of data supplied by the smaller public agencies, like those of states or cities, and by many private agencies. The best rule is to insist, as far as possible, on knowing what was done by the collecting agency at each step of the data-gathering process, from definitions to field work to final tabulation; and on noting what checks they have applied to test the accuracy reliability, and validity of their data. Only when the investigator is reasonably satisfied after a painstaking scrutiny of this kind that the data are appropriately defined and sufficiently accurate for his purpose is he justified in going forward with the work of analyzing and interpreting them. Research workers have wasted months of effort and thousands of dollars before they discovered that the material on which they were basing their conclusions was hopelessly inaccurate to start with. Obviously, no amount of mathematical treatment can make amends for data of this kind.

4. Primary Statistical Data.—The usual method of gathering firsthand data in sociological research is by means of the *schedule* or of the *questionnaire*. Both are sets of questions to be answered in blank spaces provided. The questionnaire is mailed out to informants and is not often to be recommended. Not only are the persons addressed likely to misunderstand or interpret in diverse ways the questions asked, but they seldom answer all of the questions, and many of them make no returns at all, thereby tending to produce a biased sample. A much sounder plan is to have trained interviewers with a schedule visit the persons who are to give the information,[1] or transfer the data to the schedule from available records. The procedure properly begins with the formulation of the problem, and ends with the analysis of the data, because one step logically determines another, and a given investigation should be developed as an organic whole.

5. The Schedule.—After the problem of fact finding or hypothesis testing and the general approach to it have been tentatively determined, the next step is normally to prepare the schedule. The schedule is nothing more than a list of the questions which it seems necessary to answer in order to test the hypothesis or hypotheses, or to get the facts at which the investigation is aimed. Much skill and labor are required to include all the essential questions and nothing more. Anything that is obvious or beside the point should be omitted. In addition, each question must be simple and clear, and must be answerable in terms of countable or counted units; and the same question should have approximately the same meaning for each informant. The units must be capable of objective definition, so that there will be no serious amount of disagreement about specific instances. Birth rates, an index of business conditions, marriage rates, age in years or months, I.Q.'s, "male," "female," "yes," "no," dollars, number of persons in family, occupation, and so on, are acceptable units when carefully defined in context. So much difficulty has been experienced with a term like "occupation," however, that the census bureau has prepared a large manual with a detailed list of almost every

---

[1] It is possible to mail out questionnaires to carefully stratified classes of the population, and to correct the replies in the light of answers obtained by personal visitation of much smaller samples from each stratum.

conceivable occupation, showing its schematic relationship to more inclusive occupational categories.

THE ENUMERATIVE CHECK SCHEDULE

SPECIMEN FORM EC-1 AND INSTRUCTIONS

Printed below and slightly reduced from its actual size, is a specimen copy of EC-1 with entries made to illustrate typical situations. For the persons enumerated, these include: a fully employed head of family, a housewife, a part-time worker, a worker temporarily unemployed, an unemployed worker, a new worker, a full-time student, a retired invalid, and a worker on a special Government or emergency project. The specimen EC-1 Form as set out is followed by a narrative describing the manner in which an enumerator might receive the answers which are recorded on it. The instructions printed on the back of EC-1 are reproduced on the opposite page.

Form EC-1

NATIONAL UNEMPLOYMENT CENSUS

CONFIDENTIAL

Enumeration of Persons Residing on Selected Postal Routes

A. Location of residence. 2102 North Lake St.
(Street and number, or rural or star route Lot number)
Madison
(City, town, or village)
Dane
(County)
Wisconsin
(State)

Madison Wisconsin
(Post office) (State)
University Station
(Post office station or branch)
City delivery route No. 28 Rural route No.
Village delivery route No. Star route No.

B. Does this household live on a farm? No
(Yes or no)

C. Total number of persons in this household ... 11 ... D Number less than 14 years of age ..... 2 ....

Line Number	NAME of each person 14 years of age or over living in this household (Enter last name first, then the given name)	Sex M or F / W, Neg. or O.	Color or race	Age at last birthday	Was this person working for pay (or profit) during the week of Nov. 14-20 1937? (Yes or no)	IF WORKING FOR PAY during week of November 14-20— Was he working full time? (Yes or no)	If working part time— How many hours did he work during the week?	Did he want more work? (Yes or no)	IF NOT WORKING FOR PAY during week of November 14-20— Does he usually work for pay (or profit)? (Yes or no)	Did he want work? (Yes or no)	If he wanted work Was he able to work? (Yes or no)	Was he actively seeking work? (Yes or no)	Was this person employed on WPA, NYA, CCC, or other emergency project during the week of Nov. 14-20? (Yes or no)
Column No →		1	2	3	4	5	6	7	8	9	10	11	12
1	Johnson, Philip	M.	W.	56	Yes	Yes	---	---	---	---	---	---	No
2	—— , Martha	F	W	54	No	--	--	--	No	No	--	--	No
3	—— , George	M.	W.	32	Yes	No	16	Yes	---	---	---	---	No
4	—— , Helen	F	W	28	No	--	--	--	Yes	No	--	--	No
5	—— , Arthur	M	W.	24	No	---	---	---	Yes	Yes	Yes	Yes	No
6	—— , Peter	M.	W	20	No	--	--	--	No	Yes	Yes	Yes	No
7	—— , Mary	F	W.	17	No	---	---	---	No	No	---	---	No
8	Smith , Paul	M.	W	80	No	--	--	--	No	No	--	--	No
9	Jones , Robert	M.	W.	24	Yes	Yes	---	---	---	---	---	---	Yes
10													
	Do not use this line												

IMPORTANT

1. Read carefully the instructions on other side of this form.
2. Make entries for ALL listed persons in columns 1, 2, 3, 4, and 12.
3. If answer in column 4 is "Yes," ask questions 5, 6 or questions 6 and 7 if working part time; put dashes in columns 8-11, inclusive.
4. If answer in column 4 is "No," put dashes in columns 5, 6, and 7, and ask questions in columns 8-11 as instructed.

Calls	Date	Information obtained from (to whom not completed)	*Signature of enumerator
1st	11-29	Mrs. Martha Johnson	Thomas E. Brown
2d			
3d			

Thomas E. Brown, the enumerator, begins his work on Monday, November 29, 1937. He has been instructed by the postmaster and furnished with a package of EC-1 Forms preaddressed for each dwelling on the route to which he is assigned, as well as a supply of EC-2 notices. The first EC-1 Form bears the address 2102 North Lake Street. It is not a farm, so he writes "No" in answer to "B". Mrs. Johnson answers the bell, and when Mr. Brown has introduced himself, explaining the purpose of his call, she gives him the following information about the members of her household.

There are 11 in all, including 2 not yet 14 years old. Mr. Brown writes in "11" and "2" for "C" and "D", respectively, and proceeds to list the names of the 9 grown-ups, and then to fill in the answers for each as Mrs. Johnson responds to the questions.

The head of the house is Philip Johnson, age 56. He was fully employed during the week of November 14-20 at a regular job. She is his wife, has always kept house for the family and does not want work for pay. Their oldest son, George, is 32. He has a regular job but was put on a part-time basis in September and worked only 16 hours during the week of November 14-20. Yes, he wants more work. Helen who is 28 has a job. She was out sick the week of November 14-20, but has since returned to work. Arthur, age 24, worked for several years up to last summer when he was laid off. He wanted a job during the week of November 14-20 and has been temporarily away from home 'in another city trying to find one. Peter, age 20, has not worked before but he, too, is looking for a job. Mary, age 17, is still in school. Paul Smith, age 80, is Mrs. Johnson's father who lives with her. He gave up working several years ago when his health made it impossible for him to carry on. Robert Jones, 24, is a roomer who is a laborer on a WPA project.

Mrs. Johnson also says that another family, the Smiths, live upstairs in the house. As he does not have an addressed EC-1 Form for the Smiths, Mr. Brown fills in a blank form for them and proceeds to his interview with Mrs. Smith.

A good rule is that each question in the schedule should be answerable either in terms of some standard unit like dollars and number of members in family (as defined), or in terms of a check

mark, code number, or letter that refers to a specific list. For example, after the interviewer learns the subject's occupation he may enter in the schedule the code number of the appropriate classification in the census manual of occupations. Open questions, in answer to which any word or phrase may be inserted, should be avoided. Thus the question, "To what social organizations does he belong?" is usually less desirable than a comprehensive list of social organizations to be checked, including the catchall "Others," to cover any institutions that may have been omitted from the list. The ability of informants or records to furnish sufficiently accurate answers should be considered. Questions that call for more information than is likely to be available, that rely too much on memory or on memory of the distant past, that cause fatigue, or that excite bias or involve personal interests, either are to be avoided or special provisions are to be made to estimate, overcome, or correct for the resulting errors. Questions addressed to an informant should also be inspected to see if they suggest their own answers (*e.g.*, "Do you dislike to go to school?"). The schedule should not modify the behavior it is intended to measure. Special care should be taken that the schedule is not so long as to weary or disgust the informants. If it has to be long, more than one interview should be allowed, and the informant should be paid or otherwise made to feel that the time given to it is worth his while.

On page 38 is the enumerative check schedule used as a part of the National Unemployment Census of 1937. Its purpose was not to test any hypothesis, but merely to check the number of unemployed persons enumerated by the voluntary registration plan. It meets all the requirements mentioned above, except that it employs a number of questions such as "Does he usually work for pay?" the clarity and meaning of which are not obvious.

The Enumerative Check Schedule

INSTRUCTIONS

Household Information

A. *Location.*—Give address fully, including apartment number, floor number, rear, alley, etc., if necessary to identify the household.

B. *Does this household live on a farm?*—Consider as a farm any tract of land locally so regarded.

C. Total number of persons in this household.—Include all persons living in the same household unit, including servants and lodgers, also children and others temporarily away from this household.

D. Number less than 14 years of age.—Enter total number of persons in this household who are less than 14 years of age.

Questions About Each Person

Name.—Before making the entries in any other column, list the names of all persons 14 years of age and over, then check with items "C" and "D," above, to account for every person in the household.

Write each name on a numbered line; never crowd additional names between lines or at bottom of form. For households with more than ten members 14 years of age and over, continue the listing on a second form, repeating the address.

Column 1. Sex.—Enter "M" for male and "F" for female.

Column 2. Color or race.—Enter "W" for white, "Neg" for Negro, and "O" for other. Enter persons of Mexican parentage as "white" (W). The "other" (O) group includes Indians, Chinese, etc.

Column 3. Age at last birthday.—If the exact age is not known, enter the approximate age.

Column 4. Was this person working for pay (or profit) during the week of November 14 to 20, 1937?—Enter "Yes" for each person who worked for pay (salary, wages, fees, commission, supplies, living quarters, etc.) or who worked for profit (in his own business, store, or on his own farm) at any time during the week of November 14–20. Enter "Yes" for each part-time worker, even though he worked only a few hours each day, or only a few days of that week.

Enter "No" for each person who was NOT working for pay or profit, as defined above, at any time during that week. In addition to persons who were totally unemployed, "No" should be entered for the following classes of persons:

a. Housewives and other unpaid persons engaged only in housework or helping without pay in a family business or store or on the family farm.

b. Sons, daughters, or other relatives who, without pay, help some member of the household in his work for pay or profit.

c. Full-time students, and retired or disabled persons.

d. Persons who had jobs but who were temporarily absent from work during the entire week because of sickness, strike, vacation, or other similar reasons.

.   .   .   .

6. The Instructions.—To deal adequately with the definition of the terms used in a schedule, it is customary to accompany the schedule with a set of instructions, like those that follow the check schedule of the National Unemployment Census on page 39. A reading of these instructions will give an idea of the extent to which they may improve the accuracy of the

returns. In work of this kind there is, of course, always a practical limit beyond which the matter of definition cannot be carried.

7. The Tables.—It is usually impossible to set up a schedule with much confidence unless tables to receive the returns are made up at the same time. Just what summary statistics are wanted should be listed (*e.g.*, means, proportions, correlation coefficients), and the tables needed to compute and exhibit them drawn up, together with a *transcription sheet* or cards to which all of the data will be transferred from the schedules.

Three of the many tables that were used in connection with the enumerative check schedule of the National Unemployment Census are shown below.

TABLE 2.—PERSONS ENUMERATED IN CHECK AREAS AS PARTLY UNEMPLOYED OR AS PART-TIME WORKERS, BY SEX AND HOURS WORKED DURING THE WEEK OF NOV. 14–20, 1937*
(Data for persons 15–74 years of age)

	Partly unemployed			Part-time workers		
	Total	Male	Female	Total	Male	Female
Total.....................	84,919	60,944	23,975	20,895	11,986	8,909
Reporting..............	82,898	59,438	23,460	12,388	6,538	5,580
None................	105	72	33	23	14	9
1–8 hours............	8,268	4,848	3,420	1,193	434	759
9–16 hours...........	20,499	13,899	6,550	2,636	1,211	1,425
17–24 hours..........	30,195	22,137	8,058	3,747	1,982	1,765
25–32 hours..........	18,120	14,028	4,092	3,099	1,808	1,291
33–40 hours..........	4,896	3,813	1,083	1,303	849	454
41 hours or more......	865	641	224	387	240	147
Not reporting...........	2,021	1,506	515	8,507	5,448	3,059
Per cent reporting........	100.0	100.0	100.0	100.0	100.0	100.0
None................	0.1	0.1	0.1	0.2	0.2	0.2
1–8 hours............	10.0	8.2	14.6	9.6	6.6	13.0
9–16 hours...........	24.7	23.4	27.9	21.3	18.5	24.4
17–24 hours..........	36.4	37.2	34.3	30.2	30.3	30.2
25–32 hours..........	21.9	23.6	17.4	25.0	27.7	22.1
33–40 hours..........	5.9	6.4	4.6	10.5	13.0	7.8
41 hours or more......	1.0	1.1	1.0	3.1	3.7	2.5
Median...:.............	19.3	19.9	17.7	21.0	22.5	19.3

* From DEDRICK and HANSEN, *Final Report on Total and Partial Unemployment*, 1937, Vol. IV, p. 31, The Enumerative Check Census, Census of Partial Employment, Unemployment, and Occupations, United States Government Printing Office, Washington, 1938.

TABLE 3.—PERSONS ENUMERATED IN CHECK AREAS AS NOT AVAILABLE
FOR EMPLOYMENT, BY SEX, USUAL WORK STATUS, DESIRE FOR
WORK, AND ABILITY TO WORK*
(Data for persons 15–74 years of age. Percentage not shown where less
than 0.1)

	Both sexes		Male		Female	
	Number	Per cent of population	Number	Per cent of population	Number	Per cent of population
Total not available for employment..............	608,460	41.5	102,991	14.2	505,469	68.3
Wanting but not actively seeking work.........	21,108	1.4	9,222	1.3	11,886	1.6
Usually work...........	14,082	1.0	7,491	1.0	6,591	0.9
Do not usually work.....	7,026	0.5	1,731	0.2	5,295	0.7
Wanting but unable to work................	3,471	0.2	2,264	0.3	1,207	0.2
Usually work...........	2,668	0.2	1,868	0.3	800	0.1
Do not usually work.....	803		396		407	
Not wanting and do not usually work...........	583,881	39.8	91,505	12.6	492,376	66.5

* From DEDRICK and HANSEN, *Final Report on Total and Partial Unemployment*, 1937,
Vol. IV, p. 33, The Enumerative Check Census, Census of Partial Employment, Unemployment, and Occupations, United States Government Printing Office, Washington, 1938.

As a result of constructing specific tables, the original schedule
is likely to be considerably amended and improved, especially
if a complete set of tables is made covering every important step
in the treatment to which the data are to be subjected, including
all work tables for the statistical analysis.

8. Testing the Schedule.—After a schedule has been tentatively constructed, it should be tested for accuracy, reliability,
and, if necessary, for validity. This applies to each question
separately and to the schedule as a whole.

Accuracy may be checked by applying the schedule to known
data, and noting how closely the returns agree with the a priori
information. The interviewer employed should have no prior
knowledge of the data, and should not be aware that a check is
being made. It is also sometimes possible to include in the
schedule pairs of questions that get the same information in

independent ways; but this is usually confined to a few of the most important but least reliable questions.

TABLE 4.—GAINFUL WORKERS, 1930, AND PERSONS EMPLOYED OR AVAILABLE FOR EMPLOYMENT IN ENUMERATIVE CHECK AREAS, 1937, BY SEX AND RACE AS PERCENTAGE OF POPULATION*
(Data for persons 15–74 years of age)

Year	Both Sexes			Male			Female		
	All races	White	Negro and other races	All races	White	Negro and other races	All races	White	Negro and other races
Gainful workers, 1930†	57	56	66	87	87	91	25	23	41
Employed or available for employment, 1937..............	59	58	68	86	86	87	32	30	50

* From DEDRICK and HANSEN, *Final Report on Total and Partial Unemployment*, 1937, Vol. IV, p. 35, The Enumerative Check Census, Census of Partial Employment, Unemployment, and Occupations, United States Government Printing Office, Washington, 1938.
† Data derived from Fifteenth Census of the United States, Population, Vol. V, p. 117.

Reliability is measured by trying the schedule twice on essentially the same data and comparing the results. It is often impractical to apply the schedule more than once to the same informant without introducing the memory factor or causing an undesirable response. Probably the best that can then be done is to apply the schedule to two random samples from the same universe of informants, and compare the returns. The same interviewer or interviewers should be used in each case. In all such tests, the differences observed should fall well within the range of random sampling error.[1]

A schedule, a part of the schedule, or one or more questions in the schedule, need to be tested for validity when it is not clear that they measure what is intended to be measured. This is invariably the case when broad concepts are involved. For example, if a schedule is designed to discover the number of the "unemployed" in the United States as of a certain date, it is advisable to give careful consideration to the matter of validity. Whenever a recognized and proved scale for the same purpose

[1] See Chap. XII.

already exists, all that is required is to find the amount of agreement between the returns from the two instruments, as used on the same data.   As a rule, however, this convenient situation does not occur: there is no true criterion by which to test the new instrument.

In many cases the proper approach is simply that of finding an acceptable definition.   With the help of anticipated users of the research, the investigator defines (1) what "area" of meaning of a term (*e.g.*, the "unemployed") should ideally be measured, (2) what parts of this area it is practicable to measure reliably enough for the purposes of the inquiry,[1] and (3) what parts it is not feasible to measure.   The meaning that should ideally be measured is the meaning that it is wanted to measure.   The investigator then tries to find objective and reliable indexes, which, by agreement, cover as much of the desired meaning as possible.   The remaining part that is not covered should then be clearly recognized by the investigator and his public, and both should regard the omission as not serious enough to invalidate the study.   Of course, the public may sometimes be the investigator's scientific colleagues, sometimes social welfare agencies, and sometimes the general public.   Or the investigator may merely interpret the interests of the public as he thinks best.

It will frequently happen that the persons representing the consumers of the research will differ in what they want measured.   In such a case, the choices are (1) to try to include all the desired parts of the meaning in a single index, (2) to use separate indexes for different parts of the meaning, or (3) to omit some parts of the meaning, and thereby reduce the number of people who will be satisfied with the results.

One advantage of the method of setting up an inclusive or ideal definition of the area of meaning to be measured and then marking out how much of it the given instrument can reasonably be expected to measure is the fact that it may be possible in later studies gradually to expand the area measured until consumers finally agree either that the result is a satisfactory index of the total meaning, or that the part omitted is so intangible

---

[1] *The Census of Partial Employment, Unemployment, and Occupations:* 1937, whose schedule is shown above, included persons *totally unemployed and wanting work, emergency workers* on WPA, NYA, CCC, etc., and persons *partly employed.*

and so little agreed upon that it can be disregarded. It is also better to know approximately what the instrument does and does not measure, *i.e.*, how useful it is for its purpose, than to say merely that "it measures what it measures!"

If the method of cooperative definition has not been used, or if its results are not entirely satisfactory, the question still remains whether the index measures what it is wanted to measure. Even in the more objective and simple instances, this is not always certain. Thus, if we are trying with Thorndike to measure the desirability of cities as places of residence,[1] and include urban death rates as an objective index covering one aspect of the concept of desirability, we shall need to ask if the rates have been standardized for differences in the age and sex composition of the city populations, if the out-of-town deaths occurring in local hospitals have been omitted, and so on, before we can be sure that the rates reflect differences in the incidence of fatal diseases and accidents between cities. In cases like this, the validity may be taken as established when our several questions are properly answered. But in dealing with less objective traits, this may not be enough. Suppose we include an attempt to measure the subjective trait of "friendliness" as a further element in the desirability of cities as places of residence. By the method of the cooperative definition outlined above, we may arrive at a combination of the average number of social visits and the percentage of the population belonging to social organizations as a tangible index of this subjective quality. A potential consumer of the investigation who has been consulted, however, may say that he has lived in several cities and found the people in some much "colder" to newcomers than in others, and he doubts that the index will show this difference. If the consumer from personal experience can classify certain cities as "colder to newcomers" than others, we can apply our index of friendliness and see where it places them. If the results are in agreement with his observation, he is likely to accept the index. Of course, in such cases, the experience or opinion of a single individual is not enough. We should actually need to have many persons, representative of our public, rate or score a group of cities in regard to "coldness to strangers," and compare their

[1] E. L. Thorndike, *Your City*, Harcourt, Brace and Company, Inc., New York, 1939.

ratings or scores with the results of our index of friendliness. In doing this, we should be careful to choose as raters individuals who are well acquainted through actual residence with at least some of the cities in question.

Moreover, the ratings when repeated by the same or like groups should give essentially the same results. If the several raters show little agreement among themselves, as may happen, no criterion at all will result from this procedure. In that case, we may need to face the problem of the average. Probably a certain city was actually "cold" in its treatment of some of the raters and not of others. We might then have to devise a reliable score that would reflect the proportion of the raters who regarded the city as "cold," or the amount of "coldness" that they experienced there on the average, and relate it to our index. Or we might feel it advisable to stratify our raters by socio-economic classes (*e.g.*, rich, average, poor), and get separate ratings from each class. The latter plan would require us to deal with the whole problem of the desirability of cities as places of residence from the point of view of each social class separately, which should provide a set of indexes of more value than any single index representing a gross average for all classes. In addition to subjective ratings, we might also set up, preferably by agreement, certain objective criteria of friendly or unfriendly cities, such as their methods of dealing with unfortunates, that are not included in our index, and test the latter against them.

The final test of such an index, of course, is whether in practice it proves more *useful* than other methods in selecting cities that people will actually find desirable or undesirable places of residence, in accordance with the prediction of the score card.

Ingenious ideas can often be used in testing the validity of an index. For example, if we are measuring attitude toward religion, we might see if our scale will place a group of ministers at the favorable end, a group of atheists at the unfavorable end, and average citizens for the most part in the middle. In work of this sort there are, however, many pitfalls that can be learned only from experience.

Such tests of accuracy, reliability, and validity as mentioned above imply that the schedule will be tried out in the field on a small scale and carefully revised in the light of the results before it, the instructions, and the tables are put in final form. This

preliminary trial almost invariably leads to some important changes, and should rarely be omitted from the routine of statistical research.

9. The Interviewer.—After the schedule has been carefully prepared and tested, the purpose of the interviewer or data taker is merely to see that the questions are understood and answered to the best of the ability of the informant, or that the right data are accurately copied from the proper sources. The less the interviewer says or does beyond this, the more dependable the returns should be. He must be especially careful not to suggest answers to the informant, or to bias him in any way. While this may seem to be a negative role, it is one that calls for skill and judgment. The ability to induce informants of various kinds cheerfully to give accurate information, or to extract data without error from complex or confused records, is not common.

It is often desirable to test the results obtained by each interviewer by noting whether an interviewer's returns differ too much from those of others reporting similar data. Also, when the schedules are edited, certain kinds of errors made by the interviewers may be noted. The interviewers may then be cautioned, or their work may be corrected for the personal equation.

In the gathering of information by schedule, several interviewers or clerks may be supervised by a foreman, or the investigator may do all this work himself. In any case, the investigator should participate in the actual field or library work at least enough to acquire a firsthand knowledge of the conditions under which the data were obtained, and a "feeling" for the data, as it is termed. Many an investigation has been saved or lost by the presence or absence of the analyst during the data-collecting process.

10. Editing the Schedules.—The schedules filled out during each day on a large study are generally sent in to a group of editors at the headquarters of the study. Under the direction of a chief, these clerical workers look for unfilled spaces, for inconsistent answers, and the like, on each schedule. Where necessary, a defective schedule is returned to the field or library foreman, who in turn hands it to the interviewer or the clerk whose initials appear on it. In small studies, the schedules taken during the day are often edited by the interviewers themselves each night.

11. Tabulation of the Data.—Edited schedules go to tabulators, who tally the data from the schedules to the tables, or, if machine methods are used, punch them on cards according to a code arranged for the purpose.   Machines are, of course, faster and more economical for large-scale tabulation.

The chief electrical machines now in use are the card punch and verifier, the sorting machine, and the tabulating machine. A general idea of what each does may be obtained from the following description:

The first step in the use of a card for a particular record is the designation of groups of columns as "fields."   Each field defines a section of the card in which one particular type of information will always appear.

The illustration following (Fig. 3) shows an 80-column card partly drawn up into fields.   Each field is assigned a sufficient number of columns to include the largest number of digits which it will be called upon to accommodate.
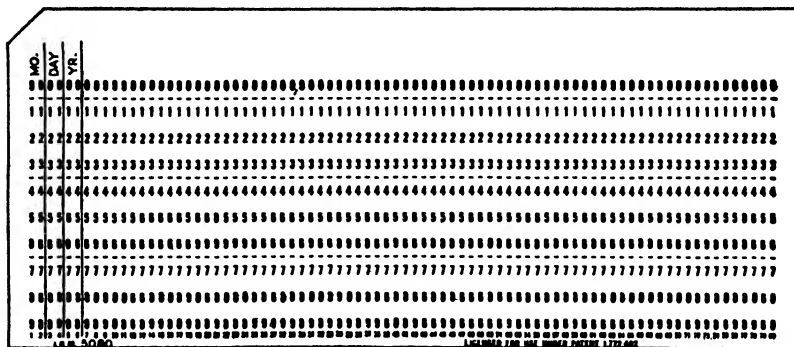
Fig. 3.—Eighty-column tabulating card.

For instance, the greatest number of months is 12 (a two-digit number), therefore, two columns are sufficient for recording this information.   The greatest number of days in a month is thirty-one, thus this field too requires only two columns.   The year is indicated by the last two digits, making two more columns necessary, etc.

Figure 4 illustrates (a 45-column) card completely laid out for a specific job, in this instance a complex (criminological) study.

At this point it is obvious that all pertinent information must be registered in the card in the form of punched holes.   The perforation of these holes is a simple matter.   The digits of the numbers to be transcribed correspond to the digits printed on the card.   Thus, to show the date Oct. 15, 1934, on the card illustrated above (Fig. 3),

the card is perforated as follows: 10–15–34. Descriptive information, such as the names of persons or products, is generally coded numerically.

Tabulating cards are perforated by means of an electric punching machine. The punch designed for the numerical system has a keyboard consisting of twelve keys, one for each punching position of a column. As a key is depressed a hole is cut and the card advanced automatically to the next column to be punched. The automatic features of the machine and the simplicity of the keyboard make the transcription of written data into punched-hole form easy, rapid and efficient.

Fig. 4.—Forty-five column card with field headings.

When punching has been completed, the cards are usually in miscellaneous order. The next step is to arrange them in sequence by some desired classification—that is, to group them according to some information which is punched in them. The Card-operated Sorting Machine is used for this purpose.

The operation of the Electric Sorting Machine is based on the position of the punched hole in a vertical column of the card. As the cards pass through the machine a brush contact is made through the hole, causing an electrical circuit to be closed. This momentary circuit causes the card to be directed to a receiving pocket which corresponds to the position of the punched hole. For example, a card punched "9" in the column under consideration is directed to the 9 pocket, a card punched "6" in the same column is directed to the 6 pocket, etc. . . . .

The automatic sort is made on one column at a time. It is apparent, therefore, that to arrange a group of cards in numerical sequence according to the data punched in a three-column field, the group is passed through the sorting machine three times. The sort is made first on the

units column, then on the tens column and finally on the hundreds column. The Card-operated Sorting Machine is entirely automatic and operates at a speed of 400 cards per minute.

The third step in the Punched Card method is the automatic compilation of the data into printed reports. This is accomplished by the Electric Tabulating Machine which is a combined adding, subtracting and printing machine. Punched cards passing through this machine actuate the various adding counters and printing mechanisms—again by means of electrical contacts . . . The machine is entirely automatic, operates at a speed of 150 cards per minute . . . [1]

12. Analysis of the Data.—The analysis of the data should proceed along the lines laid down in planning the study, although any minor modifications or extensions that later appear advisable may be made. This means that the data have already been put in work tables for computing means, percentages, standard deviations, correlations, or whatever other statistics are needed for simplifying and interpreting the findings. After these statistics have been worked out and their accuracy has been carefully checked, the investigator should state the results as simply, briefly, and clearly as he can. Only a few of the most vital tables should be presented with the text of the report, all others that seem desirable being placed in an appendix. Where graphic devices promise to be effective, they can be introduced.

Perhaps the most important things to keep in mind at this crucial stage of an investigation are to limit the conclusions to what the data show, while yet seeking to use enough imagination and insight to discover all of the pertinent information that may be extracted from the findings. There are, of course, no rules by which this can be done. Everything depends upon the ability, integrity, training, and persistence of the analyst.

13. The Amount of Error of Observation or Record in Statistical Results.[2]—The readers of sociological studies are not unreasonable when they express an attitude of skepticism toward the elaborate precision of some of the statistical techniques that are frequently applied to social data of doubtful character.

[1] HERBERT ARKIN, in G. W. Baehne, ed., *Practical Applications of the Punched Card Method in Colleges and Universities*, pp. 4–8, Columbia University Press, New York, 1935.

[2] Adapted from a paper by T. C. McCormick, On the Amount of Error in Sociological Data, *American Sociological Review*, Vol. 3, pp. 328–332, 1938.

The major difficulties involved in the estimation of errors of observation are practical rather than mathematical and theoretical in nature. Determination of the accuracy of findings is first a question of funds and time, and is tied up with administrative policies.

In England, the importance of estimating errors of record in sociological results has been recognized by Arthur L. Bowley, who writes:

If we do not know of the existence of biassed errors, which in reality pervade our estimates, there is no remedy; if we know them, we are likely to obtain more accuracy by the most erroneous corrections for them than by neglecting them. . . . In the nature of things, when we are dealing with errors we do not know their magnitude; the most we can know is their probable and possible extent. We might estimate, for instance, the percentage of unemployed in a certain year as 4.5, and add, from information in our possession (coming from a study of wage bills or the reports of relief agencies), that we considered this to be within .5 of the fact; we should then write the number 4.5 + .5, meaning that the error in the estimate as defined above was unlikely to be more than .5/4.5 = $\frac{1}{9}$, or 11 per cent, the corresponding absolute error being .5. In such a case we can also give definite limits. The percentage employed must lie between 0 and 100; and if we could actually enumerate 1 per cent of the working-class as out of work, and also 92 per cent as in work, we should know that the number required was between 1.0 and 8.0 per cent, and the maximum error in our estimate, 4.5, was 3.5/4.5 = $\frac{7}{9}$, or 78 per cent. Even this is more precise than the original statement, "the percentage is 4.5, error unknown." By further investigation we might perhaps bring the limits of error nearer to each other, and decide that it was practically certain that the percentage required was between 3.5 and 4.5; then we ought to say "the number unemployed is .04 . . . of the working class, the estimate being correct to the last figure given." This statement is of the same nature as, "The body weighs 15 lb. 3 oz., correct to an ounce."[1]

As yet, most of the theory underlying the subject of errors consists of a number of precautions that simply need to be borne in mind and observed. What seem to be the outstanding points are briefly summarized below.

[1] A. L. BOWLEY, *Elements of Statistics*, 6th ed., pp. 180, 181, 192, Charles Scribner's Sons, New York, for P. S. King & Son, Ltd., London, 1937.

(1) By definition, "The relative error in an estimate is the ratio of the difference between the estimate and the true value, to the estimate."

(2) Where the necessary a priori information exists, the results of an investigation may be compared with expectation, and the extent of the error suggested in this way. The basis of the expectation must, of course, be justified.

(3) In the absence of adequate comparative data, the only possible method of finding errors of measurement or record is to repeat the measurements, or a sufficient proportion of them. These check measurements may be made with the same measuring instruments, or by other devices and approaches, to reveal possible errors due to a particular method or scale. A change of personnel to find the amount of error attributable to the "personal equation" is also important.

(4) Where differences between the original and the check measurements are found, investigation should continue until it is possible to correct the error sufficiently for the purpose in hand by averaging or other estimate.

(5) There are two well-known kinds of error of measurement or record, whose treatment is different:

　a. Unbiased or compensating errors. Some errors occur in opposite directions, and so wholly or partly cancel out in sums, averages, and other statistics. Such random errors, however, increase the value of the standard deviation and attenuate the correlation coefficient.[1]

　b. Biased errors, or errors in the same direction:

　　(*a*) Constant error. An error that remains the same from one measurement to the other, as when a foot rule is inaccurately divided, is usually hard to detect, but very common. In social investigation it may be due to wishful thinking, to loose definition, to falsification on the part of the subjects interviewed, and so on.

　　(*b*) Accumulative error. Some biased errors increase from measurement to measurement, as when one

[1] See Chaps. VIII and X for definition of these terms.

is dealing with more and more difficult material. Thus, in taking the census, it is less easy to get accurate answers to certain questions from Negroes than from whites.

(c) Irregular noncompensating error. When measurements vary erratically, so that they affect sums and averages in important but unpredictable ways, the error must be estimated or eliminated in each separate measurement.

Apart from ingenuity and perseverance, there is no formula for finding such errors as these. Where they are suspected but not discoverable, it may be advisable to express results in the form of ratios, since biased errors are reduced in ratios and index numbers. As Bowley puts it, "The error in a ratio is approximately the difference between the errors in its two terms. . . . "

In social investigation it is especially important to avoid misleading accuracy of statement, such as carrying calculations based on crude data to two or three decimal places. The problem of how far *not* to carry significant figures should invariably be solved on the conservative side, as when, in rough population estimates running into the millions, even the tens of thousand places are given to zeros, and the hundreds of thousands are rounded off.

The final statement of an average or other statistic should include the maximum amount by which it may reasonably be in error, expressed as a percentage of the value of the statistic, as already mentioned above. For example, given the annual church attendances per individual, 58. The error of record in this figure is estimated to be 10 per cent. These facts may be expressed in some such form as 58 ± 10%.

As Bowley warns, it sometimes takes longer to estimate the approximate amount of error in the results of a study than it does to make the study itself. If sociologists give proper attention to the accuracy of their findings, therefore, they are certain to be forced by the interests of economy of time and money to simplify their problems and to investigate the same population as often as feasible. This is true if accuracy is

regarded as a purely relative thing, which need be no greater than is required to obtain a satisfactory answer to a question in hand.

Exercises

1. What use may be made of nonquantitative methods in statistical research?

2. Make a list of the main requirements of a well-chosen statistical problem, and give illustrations of what you consider good and poor, with your reasons.

3. With which of the chief sources of secondary statistical data in the United States are you acquainted?

4. Select from the latest United States Census a few definitions that seem to you (*a*) satisfactory, (*b*) unsatisfactory, and explain why you think so.

5. What are some of the most unreliable counts in the United States Census of Population, and why?

6. Collect instances of studies in which a questionnaire was mailed out and report on the proportion and representativeness of the returns received.

7. *a.* In the statistical laboratory, propose problems on a competitive basis; and after a problem has been chosen, help design a study which your class in social statistics will carry out as a semester's project.

b. Does the problem satisfy the requirements that you listed under question 2 above?

c. Indicate by which of the methods described in Chap. II the most important traits or factors concerned in this study will be measured, and show that no more exact measurement is feasible.

d. What is the dependent variable?

e. What are the main independent variables?

f. What are the important interfering factors?

g. How will the interfering factors be controlled?

h. Is the sample adequate in size?

i. What is your assurance that it is representative?

j. Does the schedule meet the demands mentioned in this chapter? Review the points.

k. Do you have all the tables that will be needed for computation and exhibition purposes, and for interpreting the data?

l. Do your instructions leave any important terms undefined, or any procedures unexplained?

m. By what methods do you propose to test the reliability and, if necessary, the validity, of your schedule?

n. To what extent have you used the method of cooperative definition to improve the validity of your indexes?

o. Will you try to measure the error due to the personal equation of the interviewers?

p. How will you estimate the amount of error in your final results?

References

American Marketing Society, *The Technique of Marketing Research*, McGraw-Hill Book Company, Inc., New York, 1937.

BROWN, LYNDON O.: *Marketing Analysis*, The Ronald Press Company, New York, 1937.

CHAPIN, F. STUART: *Field Work and Social Research*, D. Appleton-Century Company, Inc., New York, 1920.

ELMER, M. C.: *Social Research*, Prentice-Hall, Inc., New York, 1939.

ELMER, M. C.: *Technique of Social Surveys*, Jesse Ray Miller, Los Angeles, 1927.

FRY, C. LUTHER: *The Technique of Social Investigation*, Harper & Brothers, New York, 1934.

LUNDBERG, GEORGE A., *Social Research*, Longmans, Green & Company, New York, 1929.

ODUM, HOWARD W., and KATHERINE JOCHER: *An Introduction to Social Research*, Henry Holt and Company, Inc., New York, 1929.

PALMER, VIVIEN M., *Field Studies in Sociology*, University of Chicago Press, Chicago, 1928.

YOUNG, PAULINE V., *Scientific Social Surveys and Research*, Prentice-Hall, Inc., New York, 1939.

PART II
Statistical Methods

CHAPTER V

TABULATION OF FREQUENCY DISTRIBUTIONS

1. A Problem.—Before large groups of figures of any kind can be studied and interpreted, they must be arranged, or *tabulated*, in some orderly and meaningful way.

As a first exercise in the tabulation of statistical data, let us investigate the sizes of sibling families from which the students at a given college come. A *definition* is needed. What is meant by "sizes of sibling families"? Let us say that we mean the number of brothers and sisters, including the student. The sibling family, then, is the thing to be measured, while a sibling is the unit of count or measurement. Are siblings deceased to be counted? What of siblings married and moved away? What of adopted siblings, or other children not brothers or sisters reared in the family? Always such questions of definition of the thing to be measured and of the unit of measurement arise in the beginning of a careful inquiry, statistical or otherwise, and must be settled with the purpose of the investigator in view. In the present case, let us say that deceased siblings, siblings away from home, and children adopted or reared as siblings in the family shall be included.

Assuming that we have defined the thing to be counted or measured, a sibling family, and the *unit of count or measurement*, a sibling, and that the units are equal and equivalent for our purpose, we then ask each student to tell the size of his sibling family. Let us imagine that 200 students give the following sizes of sibling families.

2	2	4	1	3	1	12	6	7	2
3	1	5	2	3	4	1	2	2	5
2	6	1	4	1	2	3	3	6	3
3	2	3	1	1	2	2	2	3	4
1	1	2	2	2	2	3	8	2	2
5	8	2	3	1	2	5	3	15	6
2	3	1	1	3	9	1	8	2	2
3	1	4	3	3	4	1	2	2	2
3	2	2	1	2	2	3	7	3	6

4	2	5	2	1	2	5	3	2	8
1	7	6	9	1	1	1	12	5	4
1	3	1	1	3	3	1	5	2	2
2	6	3	5	1	5	2	2	3	5
2	5	3	3	2	5	5	2	3	2
2	4	5	1	14	1	7	3	6	2
3	2	1	4	4	4	1	2	2	6
4	11	3	1	1	2	3	6	3	4
1	3	2	1	4	4	1	6	2	2
3	2	1	4	6	2	5	3	4	2
4	4	1	4	4	1	3	4	4	10

2. The Frequency Distribution: Discrete Variable.—We have here 200 values, varying from 1 to 15. So far, the answer to our wish to know the sizes of sibling families to which the students belong is rather confusing. The *range*, or spread between the smallest and the largest values is the clearest bit of information we have.· It extends from 1 to 15, and is therefore 14. We should also like to know how many families of each size there are. As a preliminary step to this end, it is convenient to put the items in the form of an *array*, which means merely putting them in order of size.

TABLE 5.—ARRAY OF VALUES

```
1 1 1 1 2 2 2 2 2 2 3 3 3 3 4 4 5 5 6  8
1 1 1 1 2 2 2 2 2 2 3 3 3 3 4 4 5 5 6  8·
1 1 1 1 2 2 2 2 2 2 3 3 3 4 4 4 5 6 6  9
1 1 1 1 2 2 2 2 2 2 3 3 3 4 4 4 5 6 6  9
1 1 1 1 2 2 2 2 2 3 3 3 3 4 4 4 5 6 7 10
1 1 1 1 2 2 2 2 2 3 3 3 3 4 4 4 5 6 7 11
1 1 1 1 2 2 2 2 2 3 3 3 3 4 4 5 5 6 7 12
1 1 1 1 2 2 2 2 2 3 3 3 3 4 4 5 5 6 7 12
1 1 1 1 2 2 2 2 2 3 3 3 3 4 4 5 5 6 8 14
1 1 1 2 2 2 2 2 2 3 3 3 3 4 4 5 5 6 8 15
```

As a rule, however, a better form of the array is the *frequency array*. It is obtained from the original data by setting up a consecutive series of numbers covering all the observed values (here, the sizes of sibling families, 1, 2, 3, etc.) in the left-hand column of Table 6, and tallying in the right-hand column the number of times each consecutive value (size of family) occurs.[1] The latter figures are termed *frequencies*.

[1] Tallying is commonly done by making in the proper row a sloping stroke for each item (*e.g.*, family) until four strokes are made, then drawing a stroke through them for the fifth item: /XX/ /XX/ ///. The tallies

TABLE 6.—FREQUENCY ARRAY OF VALUES

Size of Sibling Family	Students Reporting Frequencies
1	39
2	55
3	38
4	24
5	16
6	12
7	4
8	4
9	2
10	1
11	1
12	2
13	0
14	1
15	1
Total........................	200

This gives the same information as Table 5, but in a much more compact form. Table 6 also satisfies our curiosity relative to the number of students reporting each size of sibling family. We see at once that most students are members of families of three or fewer siblings. .

We shall next try lumping together into *classes*, or *class intervals*, more than one size of family, with the double purpose of showing more smoothly how the students are grouped with respect to size of family, and of more easily calculating averages[1] and other statistics from the table. Imagine combining into classes family sizes 1 and 2, 3 and 4, 5 and 6, 7 and 8, and so on. We then get Table 7.

The work of combining the frequencies should be carefully checked by repetition, and it should be noted that the total is the same as for Table 6.

---

are next counted, and a figure representing the total number of items in the row is entered in the frequency column of the table. The work of tallying should be repeated, as a check, and the total should agree with the number of original items (sibling families). When machine methods of tallying åre used, the sorting machine counts the frequency in each class, and the resulting totals are simply read off and entered in the table.

[1] But the average found from Table 7 will be less accurate than that found from Table 6.

TABLE 7.—FREQUENCY DISTRIBUTION OF DATA

Size of Sibling Family	Students Reporting (Frequencies)
1 and 2	94
3 and 4	62
5 and 6	28
7 and 8	8
9 and 10	3
11 and 12	3
13 and 14	1
15 and 16	1
Total	200

Table 7 is still more concise than Table 6 and the distribution of the frequencies is more regular. There are no classes of zero frequencies, but instead a rather steady decline in the number of cases as the size of family increases, which is what one would expect.

Tables 6 and 7 are called *simple frequency distributions,* or merely *frequency distributions,* because they show the frequency of occurrence of a set of values arranged in order of size. Table 6 was also called a frequency array because successive class values increased by single units.

In Table 7 the question arises, What is now the size of family in each class? In the first class, is the size of family the average of 1 and 2 = 1.5? This is more reasonable than to say that the size is either 1 or 2. But how can a family consist of one person and a half person? Is not this taking liberties with the data? The trouble is due to the circumstance that we are dealing with a *discrete* series, *i.e.,* a series that can take only certain values (whole numbers) and no intermediate values. Thus a sibling family may contain 1, 2, or 3 members, but not 1.3, 2.7, or 3.6 members, because people always come in wholes! In contrast to a discrete series is a *continuous* series, in which the *variable*[1] may assume any whole or decimal value whatever. The ages in years of the students in a sample represent a continuous series: 19.3, 20.4, 20.6, 21.2, 21.7, 21.9, 22.1, 22.5. While a continuous series can always be mathematically averaged without logical offense, this is not true of a discrete series. For example, the ages in years of five students are 19.3, 20.4, 21.7, 21.9, 22.1, and their

---

[1] A quality (*e.g.,* sibling family) that varies in size or amount.

average (arithmetic mean[1]) age is 21.08, which is a possible value. But if five sibling families are of sizes 1, 2, 2, 3, and 5, respectively, the mean is 2.6, which is a fictitious value. We are thus faced with the dilemma either of disregarding the logical nature of a discrete series, or of abandoning the attempt to analyze it in terms of averages and other mathematical concepts. Since the purpose of an average is to simplify and represent a series, a fractional value may serve this end in the case of a discrete series, even though it is not strictly realistic, and many valuable facts can be discovered in this way that otherwise would not appear. For these reasons, discrete series are usually thrown into frequency distributions and treated in some ways as if they were continuous.

Returning now to Table 7, we may regard the average value of the two sizes of families grouped together in each class as the *mid-point* of the class (*e.g.*, for the first class of Table 7, the mid-point is $\frac{1 + 2}{2} = 1.5$). When any item is placed in a class with other terms, it is understood that it thereupon exchanges its original value for that of the mid-point of the class. For example, when a family of 4 siblings is placed in the class 3 *and* 4 in Table 7, the 4 is thereafter treated as if it were 3.5. The mid-points of any class should, therefore, always be as close as possible to the true average of the items included in the class. From Table 6 we see that the true weighted[2] mean size of the families of 1 and 2 siblings is $\frac{(39 \times 1) + (55 \times 2)}{94} = \frac{149}{94} = 1.585$, whereas our mid-point is 1.5. This is rather close agreement, and may be satisfactory for our purposes. The mid-points of other classes may be similarly tested. From Table 6 it can be seen that the mid-point of the first class in Table 7 is too small, because there are more families of 2 than of 1; but this is somewhat offset by too large a mid-point in the next class; and so on. Where one error balances another in this way, the accuracy of the mean found from the table is improved, although the mid-points of some of the classes may not be too good. Recasting Table 7 in mid-point form, we have

---

[1] (19.3 + 20.4 + 21.7 + 21.9 + 22.1)/5 = 21.08. See Chap. VII.

[2] In the weighted mean, each value (*e.g.*, size of family) is counted as often as it occurs.

TABLE 8.—FREQUENCY DISTRIBUTION OF DATA: MID-POINT FORM

Size of sibling family, mid-point (X)	Students reporting (f)	Product (Xf)
1.5	94	141.0
3.5	62	217.0
5.5	28	154.0
7.5	8	60.0
9.5	3	28.5
11.5	3	34.5
13.5	1	13.5
15.5	1	15.5
Total................	200	664.0

If we calculate the arithmetic mean from Table 8, we may compare it with the true mean found from Table 6. To find the mean, we multiply each mid-point by its frequency, sum the products, and divide by 200. This gives for the data of Table 8 a mean of 3.32, and for Table 6 a true mean of 3.315, which in this case are nearly identical. We may, therefore, approve Table 8 as far as this test is concerned.

In the case of the data on sibling families, we need to show only the lowest and highest whole numbers that can fall within a class, because we are dealing with discrete or whole numbers. These upper and lower limits of a class are called *class limits*. We may set up the *stub*, or first column, of the frequency distribution as shown in Table 7, or, if we prefer, we may write 1–2, 3–4, 5–6, and so on. Frequency distributions are usually given in class limit rather than in mid-point form, but the latter is also common. The former is better suited for tallying, the latter for computing purposes.

3. Selection of a Class Interval.—The suggestions usually given to aid in choosing a class interval for untabulated data are

1. Note the range of the data, *i.e.*, the difference between the largest and smallest values of the variable.

2. Decide about how large the interval has to be to make a significant difference in the data. For example, a difference of less than five points in a distribution of students' grades would seem to be of no consequence, since most teachers make no attempt to grade closer than that. Indeed, 10 points may seem to some sufficiently close,

If the values already have some natural spacing, the latter should often be taken as the interval. For example, the size of farms in certain regions tends to be a multiple of 40 acres: 40, 80, 120, 160, etc.

3. Consider how many class intervals would result if the size of interval tentatively chosen in (2) above were divided into the range found in (1). As a rule, from 10 to 20 intervals are desirable, although, of course, more or fewer are permissible. Revise the size of interval suggested in (2) somewhat, if it seems advisable.

4. Make all intervals of equal size, if feasible, and avoid open end intervals when possible.

5. Decide tentatively upon the mid-points and class limits of the intervals. Unless difficulty in classification is introduced, the mid-points should be whole numbers for convenience in computing, and if they can be multiples of 5's or 10's, so much the better.

6. Tally the data in the class intervals chosen. Note whether the resulting distribution reveals a smooth trend in the frequencies from one end of the scale to the other, avoiding an irregular, broken effect. If too large an interval has been used, some points of interest relative to increase or decrease of frequencies will be concealed. If the interval is too small, the distribution will lack smoothness. It is often necessary to try tabulations by larger and smaller intervals to decide these points.

7. The accuracy of the class interval chosen for computation purposes should be tested by calculating the arithmetic mean from the table and comparing it with the true mean found from the ungrouped data or from a large random sample of the ungrouped data. To obtain a class interval that will give maximum accuracy it is often helpful to use a sliding scale device like that illustrated below (Fig. 6, applied to Fig. 5).

The application of these suggestions will be illustrated.

Below is a list of the final grades of a class in statistics:

```
80 81 87 83 94
94 85 78 82 85
85 81 87 87 65
88 81 80 75 70
73 63 77 80 88
```

$$
\begin{array}{cccc}
78 & 73 & 68 & 79 \\
68 & 83 & 70 & 83 \\
72 & 84 & 74 & 88 \\
76 & 90 & 85 & 88
\end{array}
$$

Ordinarily it is not worth while to set up a frequency distribution for 41 cases, but a small number is used here for convenience.

The first step is to arrange these values in order of size, to form a frequency *array*.

TABLE 9.—FREQUENCY ARRAY OF VALUES

Grades (X)	Frequency (f)	Grades (X)	Frequency (f)
63	1	79	1
65	1	80	3
68	2	81	3
70	2	82	1
72	1	83	3
73	2	84	1
74	1	85	4
75	1	87	3
76	1	88	4
77	1	90	1
78	2	94	2
Total............	..	..	41

The range is $94 - 63 = 31$. Intervals of less than five do not seem justified by the accuracy of the data. A natural grouping, or tendency for the grades to cluster about multiples of five, would be expected. There would be only three or four intervals of 10, which seem too few. A trial interval of five, with midpoints at 65, 70, 75, etc., is shown in Table 10. These midpoints are especially appropriate, because the clustering of the grades around them should increase the accuracy of the table for computing averages and other statistics, as illustrated above. Narrow class intervals are also generally more accurate for computation purposes than are wide ones. Like the data on sibling families, these percentage grades are given only in. whole numbers, and so may most conveniently be regarded as discrete. If 65 is taken as the mid-point of an interval of 5, evidently the lowest grade that belongs in this interval is 63 and the highest is 67, so that the five grades 63, 64, 65, 66, and

67 are included. If the width of the class interval were an even number, such as 4, instead of an odd number, such as 5,
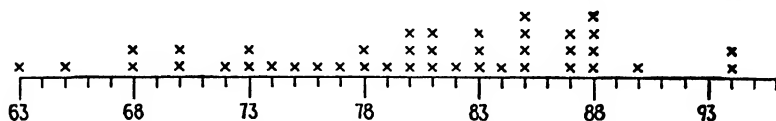
Fig. 5.—Data of array on page 66 plotted on unit scale.
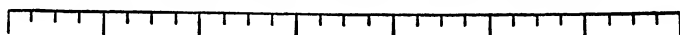
Fig. 6.—Sliding scale, with trial interval of 4.

the mid-points would be forced to take a decimal value, as was the case in Table 7.

TABLE 10.—FREQUENCY DISTRIBUTION OF STUDENTS' GRADES

Grades (X)	L*	Frequency (f)	fX
65	63–67	2	130
70	68–72	5	350
75	73–77	6	450
80	78–82	10	800
85	83–87	11	935
90	88–92	5	450
95	93–97	2	190
Total............	.....	41	3,305

* *L* means class limits.

From inspection of Fig. 5, where each value has been plotted along the grade scale, it appears that the above choice of class intervals throws the mean below the mid-point in the intervals 63–67, 68–72, 73–77, 83–87, 88–92, 93–97. Only in two intervals, however, 88–92, and 93–97, is the lack of balance serious. In one interval, 78–82, the mean is at the mid-point. The true mean of the series, computed from the separate values, is 80.195. The mean found from the frequency distribution of Table 10 is 80.610, which is 0.415 too high, as would be expected. If this amount of inaccuracy is considered important for the purpose in hand, an attempt to obtain a better class interval should be made. This may be facilitated by making a sliding scale from ordinary coordinate paper, using the same units as in Fig. 5 (see Fig. 6). Class intervals of different sizes may be measured

off on the sliding scale, and each tested in turn against the scale in Fig. 5. In the case of any given interval, the trial scale (Fig. 6) is moved along the fixed scale (Fig. 5) until the frequencies shown on the fixed scale are as evenly balanced as possible around the mid-points of the trial class interval on the sliding scale. If satisfactory, the values of the class limits may then be read off on the fixed scale from the intervals on the sliding scale when in this position of balance. Usually, some inaccuracy is inevitable in the use of class intervals. The problem is to keep it within such limits that no serious damage will result to the conclusions of the study.

TABLE 11.—BIRTH RATES PER 1,000 IN 150 APPROXIMATELY EQUAL
POPULATIONS

Class limits (1)	Mid-points (2)	Frequencies (3)
12.5–13.4	13	3
13.5–14.4	14	15
14.5–15.4	15	26
15.5–16.4	16	31
16.5–17.4	17	43
17.5–18.4	18	25
18.5–19.4	19	5
19.5–20.4	20	2
Total........	..	150

4. The Frequency Distribution: Continuous Variable.—A continuous variable, such as birth rates, is tabulated in the same way as a discrete variable, except that a slight modification is needed in finding class limits from mid-points, and vice versa. In Table 11, given the mid-points of col. (2), what are the lower and upper values of each class within which the birth rates can be classified? The boundary line between any two mid-points should evidently be halfway between them—in this case $\frac{1}{2}$ of 1, or 0.5 unit above the lower or below the higher mid-point. We thus get as our class limits in col. (1), 12.5, 13.5, 14.5, and so on. Notice that the upper limit of any class is made slightly smaller than the lower limit of the class just above, to indicate that a case falling exactly on the border line between two classes is

placed in the upper class rather than in the lower.[1]   Assuming that our original data carry only one decimal place, it is enough to write the upper limit of the class 12.5 to 13.5, for example, as 13.4; but if the data carried two decimal places, the upper limit should be written 13.49; and so on.

To find the mid-points, given the class limits, of the continuous variable of Table 11, we add the lower limit of an interval to the lower limit of the *interval next higher* on the scale, then average them.

$$\frac{12.5 + 13.5}{2} = 13$$

$$\frac{13.5 + 14.5}{2} = 14$$

and so on.

5. The Frequency Distribution: Nonquantitative Variable.[2]— Let us imagine that, instead of adopting quantitative classes for sizes of sibling families, as shown in Table 7, we had asked the students to state whether or not the size of their sibling family was large, medium, or small, without telling them what sizes of families should be placed in each of the three categories. We might then get a table something like Table 12.

TABLE 12.—SIZE OF SIBLING FAMILIES

Size of Sibling Family	Students Reporting
Small	116
Medium	46
Large	38
Total	200

We may now call attention to three requirements of *classification* that were not mentioned in our previous work, although they were tacitly assumed.   The first of these is that the categories must be *mutually exclusive*.   The second is that they must be *exhaustive*.   The third is that there must be only *one basis* of classification at a time.

---

[1] Theoretically, the frequency of a value that is identical with a class limit should perhaps be divided equally between the two classes above and below; but in most practical work the method suggested above is more convenient and sufficiently accurate.

[2] A nonquantitative variable is a quality that varies in amount, but is not measured in terms of units.

With respect to the last-named requirement, the basis of classification in Table 12 is size of sibling family. There is no evidence that any other principle was used in this table. A question may be raised, however, about the first requirement. If we checked to see, we should certainly find that some sibling families of three were listed as small and some as medium, and that similar errors were made in the case of families of other sizes. Moreover, if the sibling families reported above were entered in Table 12 by two independent investigators, even though neither happened to put families of the same size in two different classes, there is little chance that they would both classify each size of family in the same class. Some would regard a family of three as medium, others would regard it as small. Their finished tables would not show the same frequencies in each class. Because these difficulties of classification multiply with the number of classes, it is usually advisable to have very few classes in a qualitative table, *e.g.*, three in Table 12. This limits the analysis to broad categories.

Regarding the principle of exhaustiveness of classification, we need to ask: Were there any families that could not be classified in one of the three classes of Table 12? Apparently there were not, so the table passes this test.

Can we calculate the mean of Table 12, as we did in Table 8? At once the question arises, what are the values of the mid-points in Table 12? Since the classes in this table are *not quantitative*, no quantitative values can be assigned to their mid-points. We therefore discover that we are unable to analyze a nonquantitative table by the use of the mean. All that we can do is to say that the modal class, or the class containing the largest frequency, is that of small families.

From this illustration, we learn that nonquantitative tables not only are likely to violate the logical principle of classification, which requires that the several categories be mutually exclusive, but that they also do not lend themselves to the calculation of the mean and other basis statistical measures by which quantitative tables are customarily analyzed. For such reasons as these, quantitative classes are always to be preferred to qualitative for purposes of statistical analysis. The latter should be employed only where quantitative classes are not obtainable.

6. The Frequency Distribution: Table Structure.[1]—The main heading of a frequency table is called the *title;* the left-hand column with its heading, the *stub;* and the heading of the right-hand column, the *caption.* These are illustrated in Table 13.

(Title) TABLE 13.—THE SIZE OF SIBLING FAMILIES OF 200 STUDENTS OF SOCIOLOGY, BLANK COLLEGE, 1939–1940

(Stub) Siblings in Family	(Caption) Students Reporting
1– 2	94
3– 4	62
5– 6	28
7– 8	8
9–10	3
11–12	3
13–14	1
15–16	1
Total............................	200

As far as feasible, a table should be *self-explanatory,* but no unnecessary word or figure should be included. The title should usually mention the variable in the stub first; the units of the caption and their number, second; and any further subdivisions of the stub or caption. It should also generally mention the date and place. The purpose of the stub and the caption

TABLE 14.—THE SIZE OF SIBLING FAMILIES OF 200 STUDENTS, BLANK COLLEGE, 1940–1941, BY URBAN AND RURAL RESIDENCE

Siblings in family	Students reporting		
	Total	Urban	Rural
1– 2	94	28	66
3– 4	62	20	42
5– 6	28	11	17
7– 8	8	5	3
9–10	3	2	1
11–12	3	0	3
13–14	1	0	1
15–16	1	0	1
Total..........	200	66	134

[1] More detailed discussion of this topic and of tables that do not represent frequency distributions will be found in the fifth and seventh references at the end of this chapter.

is simply to indicate the nature of the entries in the columns. The "Total" row is often placed at the top instead of at the bottom of the table.   One customary type of *ruling* is shown in Table 14.   If it is desirable to block off one part of a table from another, this may be done by means of a heavy or double ruling.

The chief requirement of a good table is that it be *simple* and *clear*.   For this reason, it is generally unwise to subdivide the stub or the caption very often.   One simple subdivision of the caption is shown in Table 14.

In the case of every subclassification of the data in a table, the principles of classification already mentioned apply.

Exercises

1. Tabulate each of the two following series in a frequency distribution, showing class limits, and test the accuracy of each of the tables.   Note: The population numbers are so large that only whole hundreds or thousands should be used as class limits and mid-points; but in finding mid-points from the class limits the method suggested for a continuous variable should be used.   An interval at least as small as 5,000 seems to be needed to differentiate between the bulk of the county populations under 40,000.   But above that point increasingly large intervals are appropriate.   The last interval may be taken as "300,000 and over," with the actual population of the single largest county, 318,587, given in a footnote.   A table may be "broken" to avoid many intervals without frequencies.

Georgia Counties, 1930*

County	Population	Population per square mile	County	Population	Population per square mile
1	13,314	29.3	46	21,599	50.1
2	6,894	20.9	47	18,025	45.4
3	7,055	26.0	48	22,306	65.2
4	7,818	21.9	49	9,461	45.5
5	22,878	74.5	50	18,273	34.9
6	8,703	43.7	51	2,744	7.6
7	12,401	73.8	52	10,164	22.7
8	25,364	53.9	53	18,485	51.2
9	13,047	51.0	54	24,101	31.5
10	14,646	32.2	55	7,102	24.7
11	77,042	278.1	56	12,969	32.3
12	9,133	44.6	57	8,665	37.0
13	6,895	15.9	58	48,667	96.9
14	21,330	41.5	59	10,624	43.0
15	5,952	13.8	60	15,902	57.0
16	26,509	39.7	61	318,587	1,650.7
17	29,224	30.6	62	7,344	16.7
18	9,345	46.0	63	4,388	25.8
19	10,576	37.2	64	19,400	44.2
20	6,338	8.9	65	16,846	44.9
21	9,903	46.9	66	19,200	43.2
22	8,991	39.4	67	12,616	30.3
23	34,272	69.7	68	27,853	63.3
24	9,421	55.7	69	12,748	44.0
25	4,381	5.5	70	30,313	69.4
26	105,431	284.9	71	13,070	24.7
27	8,894	40.8	72	13,263	46.7
28	15,407	47.0	73	11,140	22.2
29	20,003	46.6	74	15,174	58.1
30	25,613	224.7	75	9,102	31.9
31	6,943	34.2	76	15,924	49.1
32	10,260	72.3	77	11,280	25.5
33	7,015	9.4	78	12,199	32.3
34	35,408	100.3	79	21,609	60.9
35	19,739	31.2	80	8,594	26.8
36	30,622	57.9	81	8,118	27.1
37	8,793	25.1	82	20,727	32.1
38	11,311	46.9	83	12,908	37.7
39	25,127	56.7	84	12,681	43.4
40	7,020	22.0	85	8,992	23.9
41	17,343	62.6	86	9,754	53.0
42	4,146	22.3	87	5,190	27.2
43	3,502	16.2	88	32,693	40.6
44	23,622	40.5	89	8,328	25.5
45	70,278	258.4	90	8,153	15.0

*From the Fifteenth Census of the United States, 1930, Bureau of the Census, Washington, D. C.

GEORGIA COUNTIES, 1930.*—(*Continued*)

County	Population	Population per square mile	County	Population	Population per square mile
91	7,847	27.0	126	20,503	25.8
92	4,180	10.6	127	7,389	30.8
93	29,994	62.1	128	23,495	112.4
94	4,927	17.6	129	11,740	70.7
95	9,014	31.4	130	11,114	27.0
96	5,763	12.3	131	26,800	58.8
97	16,643	50.1	132	8,458	27.1
98	14,921	52.5	133	6,172	29.1
99	6,968	19.4	134	15,411	33.1
100	22,437	45.2	135	10,617	31.2
101	9,076	35.9	136	14,997	40.2
102	6,730	49.1	137	18,290	51.8
103	23,620	43.1	138	32,612	61.5
104	11,606	24.7	139	16,068	66.1
105	10,020	52.7	140	17,165	43.7
106	12,488	32.0	141	4,346	24.0
107	9,215	26.9	142	7,488	28.6
108	57,558	244.9	143	36,752	84.5
109	17,290	66.0	144	11,196	48.5
110	8,082	47.0	145	8,372	26.7
111	12,927	25.6	146	6,340	19.6
112	12,327	38.0	147	19,509	61.5
113	10,268	57.4	148	26,206	60.7
114	9,687	41.9	149	21,118	63.8
115	12,522	36.3	150	26,558	34.4
116	10,853	45.8	151	11,181	27.7
117	25,141	79.3	152	25,030	37.4
118	9,005	34.9	153	12,647	20.6
119	8,367	23.2	154	5,032	16.7
120	3,820	26.5	155	9,149	34.7
121	6,331	16.8	156	6,056	24.7
122	17,174	41.7	157	20,808	73.5
123	72,990	228.8	158	13,439	33.3
124	7,247	60.9	159	15,944	34.8
125	5,347	34.7	160	10,844	23.0
			161	21,094	32.4

* From the Fifteenth Census of the United States, 1930, Bureau of the Census, Washington, D. C.

2. Subdivide the table of county populations prepared in Exercise 1 above according to population per square mile, choosing your own points of division in the latter factor.

3. Open a textbook in elementary sociology to some page at random, and classify each word on the page as "Very short," "Short," "Aver-

age," "Long," "Very Long." Show the results in tabular form. Do the same thing for an elementary textbook in economics, and compare the length of words in the two tables.

4. It is wanted to know the occupation of the fathers of students majoring in sociology. The students are asked to check the form below:

> Laborer
> Businessman
> Professional
> Farmer

Is this satisfactory?

Where would a carpenter-contractor be placed? A policeman? The proprietor of a radio repair shop?

5. A study is to be made of farm wages in your state. How would you define the unit of study?

6. Explain and illustrate the meaning of these terms: (*a*) array, (*b*) range, (*c*) frequency distribution, (*d*) class interval, (*e*) mid-point, (*f*) class limits, (*g*) grouped data.

7. What is the effect of tabulation by class intervals on the accuracy of statistics calculated from a table? Why is this?

References

CHADDOCK, R. E.: *Principles and Methods of Statistics*, Chaps. IV and V, Houghton Mifflin Company, Boston, 1925.

Fifteenth Census of the United States, 1930: *Population*, Vol. II.

GARRETT, H. E.: *Statistics in Psychology and Education*, pp. 1–8, Longmans, Green & Company, New York, 1926.

MILLS, F. C.: *Statistical Methods*, rev. ed., Chap. III, Henry Holt and Company, Inc., New York, 1938.

MUDGETT, B. D.: *Statistical Tables and Graphs*, Part I, Chap. III, Houghton Mifflin Company, Boston, 1930.

SORENSON, H.: *Statistics for Students of Psychology and Education*, pp. 16–27, McGraw-Hill Book Company, Inc., New York, 1936.

WALKER, H. M., and W. N. DUROST: *Statistical Tables, Their Structure and Use*, Parts I and III, Bureau of Publications, Teachers College, Columbia University, New York, 1936.

WHITE, R. C.: *Social Statistics*, Chap. VI, Harper & Brothers, New York, 1933.

YULE, G. U., and KENDALL, M. G.: *An Introduction to the Theory of Statistics*, Chap. VI, Charles Griffin & Company, Ltd., London, 1937.

CHAPTER VI

GRAPHS

1. Graphs of Frequency Distributions.—It is often helpful in interpreting a frequency distribution or other statistical data to show the facts in graphic form. One method of picturing a simple frequency distribution is by means of the *histogram*. Table 15 may be represented as shown in Fig. 7.

TABLE 15.—GRADES MADE BY 41 STUDENTS OF STATISTICS, BLANK COLLEGE, 1939–1940

Grades, per cent	Students	Accumulated frequency	Accumulated percentage frequency*
63–67	2	2	4.9
68–72	5	7	17.1
73–77	6	13	31.7
78–82	10	23	56.1
83–87	11	34	82.9
88–92	5	39	95.1
93–97	2	41	100.0
Total............	41		

* Each accumulated frequency is expressed as a percentage of 41, *e.g.*, $\frac{7}{41} = 17$.

In connection with the histogram, it should be noticed that, if the class intervals are taken as one unit each, the area of the figure is equal to the total frequency of the table. In Fig. 7, for example,

$$\text{area} = 2 \times 1 + 5 \times 1 + 6 \times 1 + 10 \times 1 + 11 \times 1 + \\ 5 \times 1 + 2 \times 1 = 41.$$

A second device for picturing a simple frequency distribution is the *frequency polygon*, which is constructed by connecting the mid-points of the class intervals of the histogram by straight lines. It is shown in Fig. 8. If it is extended to the base line at the mid-points of the intervals next beyond the end intervals,

76

and all equal intervals are taken as one unit in width, its total area is equal to that of the total frequency of the table, but the area over any one interval is usually not equal to the frequency in that interval.
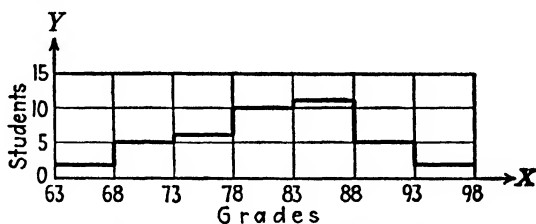
FIG. 7.—Histogram of Table 15.[1]

A histogram of the simple frequency distribution of **Table 16**, which has unequal intervals, appears in Fig. 9.

TABLE 16.—AGE DISTRIBUTION FOR MEXICANS IN THE UNITED STATES, 1930*

Age, Years	Number (Thousands)
Under 5	21.48
5–9	20.55
10–14	14.81
15–19	13.72
20–24	14.65
25–29	13.53
30–34	10.11
35–44	16.30
45–54	9.53
55–64	4.60
65–74	1.96
75 and over	0.88
Total	142.12

* Adapted from Abstract of the Fifteenth Census of the United States, 1930, Bureau of the Census.

If we let each interval of five years on the base line be one unit, then, of course, an interval of 10 years will be two units, and the height of the rectangle in a 10-year interval will be one-half of the tabular frequency in that interval. The end interval, "75

---

[1] Notice that graphs, such as those of frequency distributions, which involve two sets of measurements, are erected on the framework of two graduated straight lines drawn at right angles. The horizontal line is called the *X* axis, the perpendicular line the *Y* axis. Frequencies are conventionally measured on the *Y* axis (but see Fig. 13), scale values on the *X* axis (see Fig. 7).

and over," in Table 16 is of unspecified length, and so cannot be accurately represented geometrically. It is accordingly omitted from the graph, and its frequency removed from the total. The sum of the areas of the remaining rectangles is then equal to the corrected total frequency of the table. Moreover,
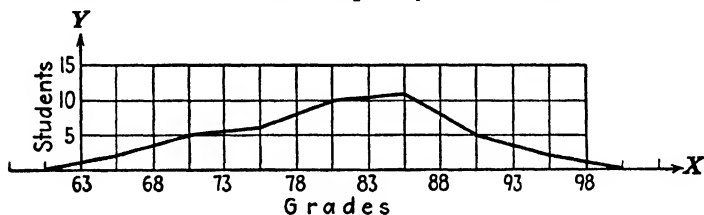
FIG. 8.—Frequency polygon of Table 15.

the area of each rectangle is equal to the frequency in the corresponding interval.

If a polygon is drawn on Fig. 9 in the usual way, neither the total area of the polygon nor the area in any interval will be equal to the corresponding tabular frequency. The total area can be made equal to the total frequency, however, if the polygon is drawn to the mid-points of five-year intervals throughout, using the same frequencies (heights) as in Fig. 9.
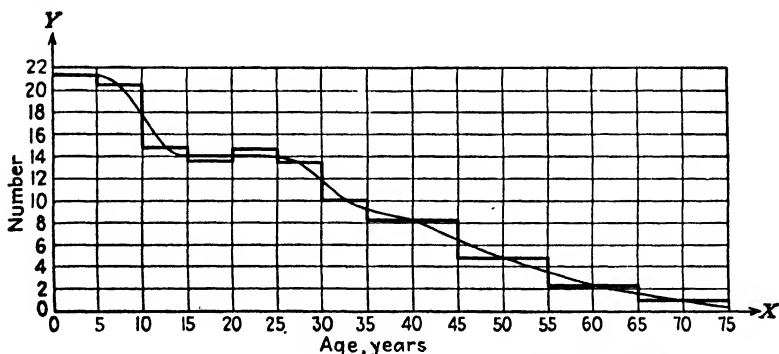
FIG. 9.—Histogram of Table 16, with unequal intervals.

Notice that if the frequencies were known and graphed for each year of age, instead of for each five- or 10-year age interval, the rectangles of the histogram in Fig. 9 would become more numerous and narrow. If then the frequencies in each year were separated by months, we should have still more and narrower rectangles. If this process of subdivision of intervals

were continued indefinitely, we should have a smooth curve instead of a histogram or a polygon in Fig. 9. It is apparent that if each minute interval were then regarded as being one unit in width, the area under any part of the smooth curve would be equal to the frequency over the same portion of the table (see Fig. 10). A great deal of use is made of this fact in some of the chapters that follow.

A polygon may be smoothed by passing through it a freehand curve. This is a somewhat questionable way of judging how the distribution would appear if the size of the sample were greatly increased.
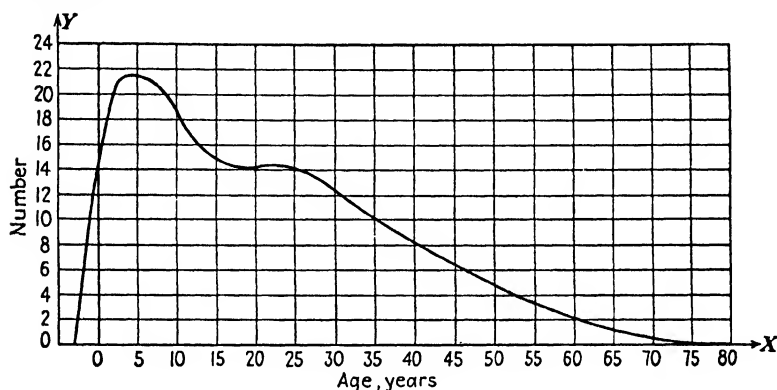
Fig. 10.—Histogram of Fig. 9 reduced to a smooth curve.

When histograms or polygons are to be compared, they should be graphed in terms of percentage rather than absolute frequencies.

A very useful type of graph in the interpretation of a frequency distribution is the *cumulative curve*, or *ogive*. The accumulated frequencies for Table 15, forming a *cumulative frequency distribution*, may be seen in the last column of that table. Since each accumulated frequency merely shows the total number of values that are less than the lower limit of the class just above on the scale, a frequency distribution in accumulative form is sometimes called a *less than frequency* distribution. Plotting should be done carefully on coordinate paper, in order that the resulting graph may be accurate enough for computing purposes. The cumulative frequency curve for Table 15 is shown in Fig. 11.

Notice in **Fig. 11** that the frequency in each class interval is plotted on the *upper class limit*, to show that a particular number of students made a grade less than the one indicated by that limit.   Thus, 34 students in the given course made grades less than 88, and 39 made grades below 93.

Not only does the cumulative frequency curve give a picture of the distribution of frequencies that is different from that shown by the histogram or polygon, but it may also be employed for interpolation and computation.   If, for example, we are given Table 15 but know nothing else about the data, and wish to change the class limits of the table, we can sometimes do this
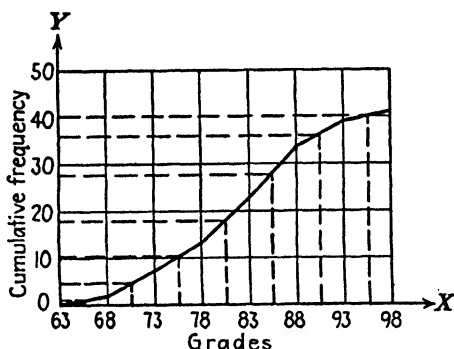
Fig. 11.—"Less-than" cumulative frequency curve or ogive, for Table 15.

most conveniently by means of the cumulative curve.   Suppose that we want the class limits of 65 to 69, 70 to 74, 75 to 79, and so on.   How many students made grades falling in each of these new intervals?   This can be decided approximately by erecting perpendiculars at the points 65, 70, 75, and so on, on the base scale, noting where they intersect the cumulative curve, and drawing horizontal lines from these points of intersection to the frequency scale at the left.   Thus, the horizontals cut the frequency scale at approximately the values 1, 5, 10, 18, 29, 37, 40.   We can accordingly set up a new frequency table, Table 17, whose last column is obtained by subtracting in the second column the accumulative frequency in each class from that in the class just above.

When it is desired simply to halve or combine the class intervals of a simple frequency distribution, the work may be done by direct division or addition more easily than by the use of a

TABLE 17.—ILLUSTRATING CHANGE OF CLASS INTERVALS OF TABLE 15, BY
USE OF CUMULATIVE FREQUENCY CURVE

Grades, per cent	Accumulative frequency	Students
60–64	1	1
65–69	5	4 .
70–74	10	5
75–79	18	8
80–84	29	11
85–89	37	8
90–94	40	3
95–99	41	1
Total..........	..	41

cumulative curve.   Thus, if in Table 15 the intervals are to be
halved, the frequencies in each interval are also halved cor-
respondingly.   It is often desirable, however, to modify this
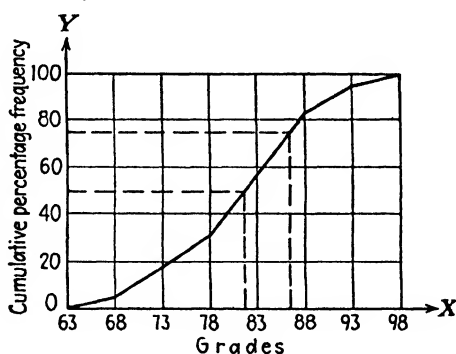
FIG. 12.—Ogive in terms of percentage frequencies.

method somewhat by allowing for the shape of the curve.   For
example, if the curve is rising in the interval, more of the fre-
quencies may be placed in the upper than in the lower subdivision
of the interval.

Percentage frequencies are often substituted for absolute
frequencies on the ogive.   Figure 12 is the same as Fig. 11 except
for this change.   From it we read on the Y axis that 50 per cent
of the students made a grade of less than 82 on the X axis,
approximately; 75 per cent make less than a grade of about 87;
and so on.   The readings can be more accurate if finely ruled
coordinate paper is used.

Values may be accumulated on both scales, X and Y, and expressed as percentages of their respective totals. This has been done in Table 18 and Fig. 13. Each pair of accumulative percentages determines a point, and they are called the *coordinates* of the point. For example, the first two accumulative percentages in the table furnish the coordinates (6.6, 0.2), the one on the left (6.6) being an X value and the one on the right (0.2) a Y value. The point is located on the chart by going a distance of 6.6 percentage units from 0 along the X axis, and then perpendicularly up a distance of $0.2Y$ percentage units.

TABLE 18.—NUMBER OF FARMS BY SIZE, KANSAS, 1930*

Size of farm, acres	Number of farms	Total acreage	Per cent		Accumulated per cent	
			Farms	Acres	Farms	Acres
Under 20...	11,004	86,739	6.6	0.2	6.6	0.2
20– 49....	9,264	312,710	5.6	0.7	12.2	0.9
50– 99....	19,226	1,475,364	11.6	3.1	23.8	4.0
100– 174....	42,920	6,319,557	25.8	13.5	49.6	17.5
175– 259....	25,481	5,565,698	15.4	11.8	65.0	29.3
260– 499....	38,385	13,796,240	23.1	29.4	88.1	58.7
500– 999....	15,055	10,243,252	9.1	21.8	97.2	80.5
1,000–4,999....	4,487	7,184,515	2.7	15.3	99.9	95.8
5,000 and over.	220	1,991,572	0.1	4.2	100.0	100.0
Total........	166,042	46,975,647	100.0	100.0		

* Adapted from Fifteenth Census of the United States, Bureau of the Census.

The resulting curve is called the *Lorenz curve*. From it we can see that 50 per cent of the farms, *i.e.*, the small farms (reading from the left on the X axis) include 18 per cent of the total farm acreage (reading from the bottom on the Y axis); that about $100 - 65 = 35$ per cent of the farms, *i.e.*, the large farms (reading from the right on the X axis) include $100 - 30 = 70$ per cent of the total farm acreage (reading from the top on the Y axis); and so on.

Further uses of the cumulative curve for computation will be shown later, under the topic of partition values (Chap. VIII).

2. Graphs of Time Series.—Statistical data often take the form of a *time series*, rather than of a frequency distribution. A time series is a set of values of a variable that correspond to

certain time intervals, such as years or months. For example, the populations of a state in 1920 and again in 1930 are a very brief time series (see Fig. 14).

In plotting the increase of one variable, *e.g.*, the population of a state, in terms of a second variable, *e.g.*, years, it is often
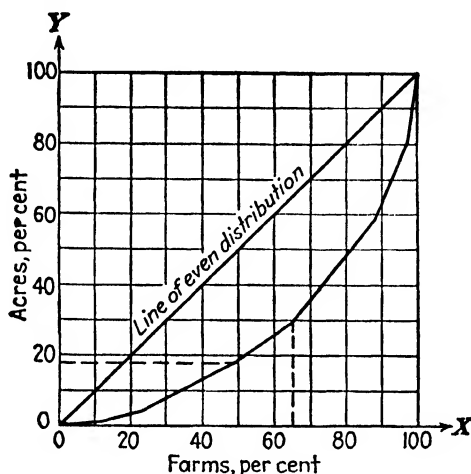
FIG. 13.—Lorenz curve, for Table 18.
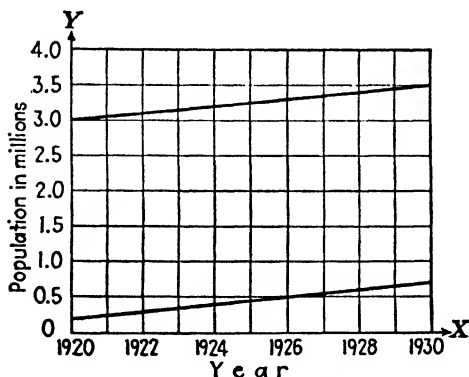
FIG. 14.—Population growth in absolute amounts.

of more interest to show the proportionate increase than the absolute increase. For example, if a population of 3.0 millions increases to 3.5 millions in 10 years, the increase is much less impressive than when a population of 0.2 million increases to

0.7 million in the same period. Yet if the absolute increase is plotted, this difference will not appear, as may be seen from Fig. 14, where the two growth lines are exactly parallel. To meet this objection, the percentage increase may be plotted. The growth from 3.0 to 3.5 millions is a percentage increase of 17, that from 0.2 to 0.7 million is a percentage increase of 250. This is shown in Fig. 15, where the line representing the growth of the population of 0.2 million is much steeper than that representing the growth of the population of 3.0 million. In making Fig. 15, the rate of growth in terms of the initial population is required. Instead of going to the trouble of computing these rates, much the same results may be accomplished by plotting
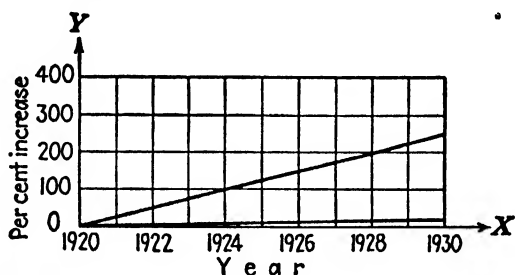
FIG. 15.—Population growth in terms of percentage increase.

the absolute figures on a semilogarithmic scale. The latter method is usually preferred to the former, because semilogarithmic paper can be obtained at small cost, and the use of it saves much labor.

Figure 16 shows the above population figures plotted directly on semilogarithmic paper.

In Fig. 16, notice that the increase in population from 0.2 to 0.7 million is again represented by a much steeper line than is the increase from 3.0 to 3.5 millions.

While the semilogarithmic scale does not show in strictly accurate proportion one to another all percentage changes, it represents equal percentage changes by equal slopes, and saves much labor compared with percentage charts such as that shown in Fig. 15.

In using semilogarithmic paper, the repeated series of values 1 to 9 usually printed on the vertical scale may be multiplied by any constant, provided the constant is applied to the whole scale.

Thus, in Fig. 16 the scale may be multiplied, say, by 7, by 0.5, or by any other number, when thereby it will be made more convenient for the plotting of particular data. A semilogarithmic scale cannot contain a zero value.
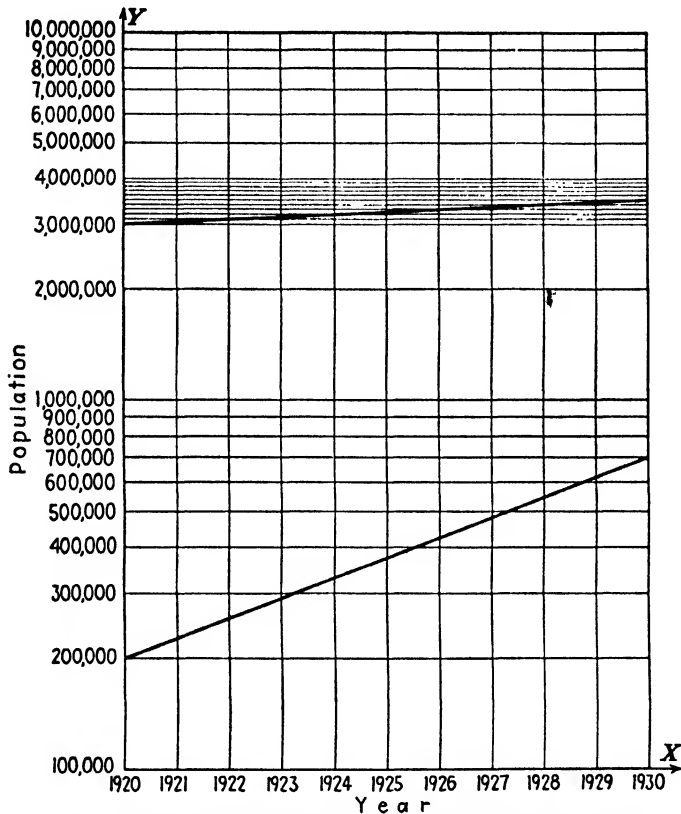
Fig. 16.—Population growth plotted on semilogarithmic paper.

In all graphic representation of data, the shape of the curve or figure is affected by the ratio of the X and Y scales. Since this ratio is usually a matter of arbitrary choice, advantage is sometimes taken of the opportunity to produce certain desired impressions. Figures 17 through 19 from Table 19 illustrate only three of many possibilities.

TABLE 19.—INCREASE IN ENROLLMENT OF THE BLANK MILITARY ACADEMY,
1928–1938

Year	Enrollment	Year	Enrollment
1928	110	1934	118
1929	113	1935	120
1930	114	1936	122
1931	113	1937	127
1932	118	1938	130
1933	119		

In Fig. 17, a rather moderate increase in enrollment is made impressive by (1) using a large single-unit spacing on the Y scale, (2) starting the increase from the base (bottom) line, and so avoiding any comparison between the amount of increase and the original volume of enrollment, (3) showing each year's increase as a percentage of the enrollment in 1928, instead of as a percentage
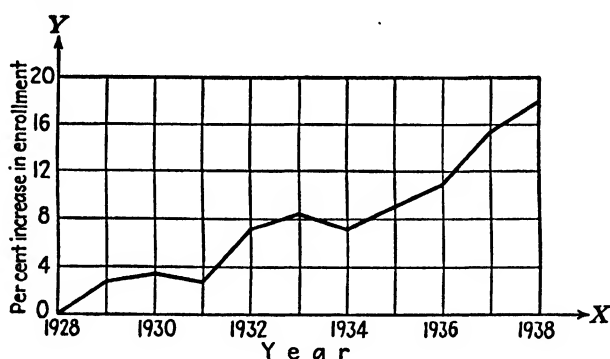
FIG. 17.—Graph of data of Table 19.

of the enrollment of the preceding year.   Figure 18 removes criticism (2) above, and avoids criticism (3) by using absolute enrollment figures instead of percentages of increase relative to the total enrollment in 1928.   Figure 18 is still open to criticism (1) above, because the ratio of the X and Y units is not changed.   In fact, the X and Y units are different in nature, so that it is impossible to say when one bears a just relation to the other.

Figure 19 meets criticism (3) by plotting the enrollment figures on a semilogarithmic scale.   The total enrollment is not

entirely pictured in the diagram because the semilogarithmic
scale begins at 1 instead of at 0, but this is a minor matter.
Evidently, the growth of the school makes a much poorer showing
in Fig. 19 than in either Fig. 17 or Fig. 18. Probably Fig. 19
gives the most realistic picture of the facts in this particular case.
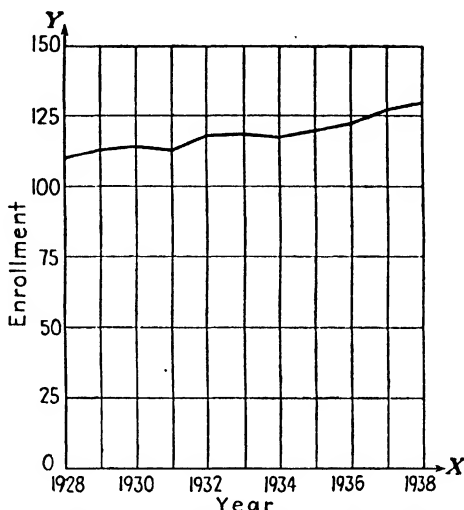
Fig. 18.—Absolute increase in enrollment of the blank military academy,
1928–1938.

3. Miscellaneous Graphs.—A common device for the graphic
comparison of amounts or percentages is the bar chart, either
upright or horizontal. The histogram of Fig. 7 above can be
regarded as essentially an upright bar chart. Figure 20 shows a
horizontal bar chart applied to Table 20.

TABLE 20.—PERCENTAGE OF FEMALES 15–44 YEARS OF AGE MARRIED,
SELECTED EUROPEAN COUNTRIES*

Country	Percentage of Females Married
Bulgaria	67.0
England and Wales	48.5
France	57.1
Germany	48.4
Italy	48.4
Sweden	42.3

* Adapted from W. S. THOMPSON, *Population Problems*, 2d ed., p. 104, McGraw-Hill Book
Company, New York, 1935.
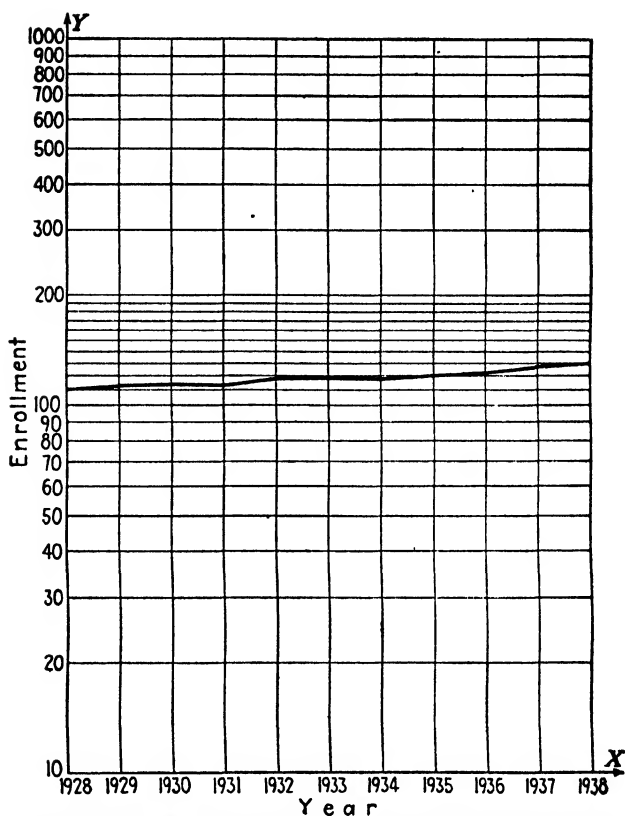
Fig. 19.—Rate of increase in enrollment of the blank military academy, 1928–1938.
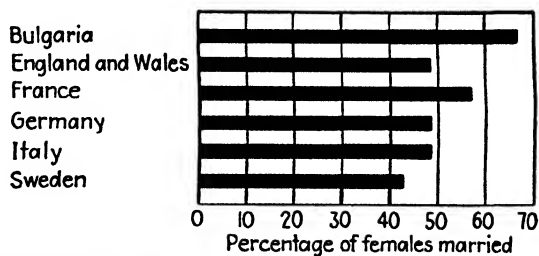
Fig. 20.—Percentages of females 15–44 years of age married in selected European countries. (*From W. S. Thompson, op. cit., p.* 104.)

Two variations of the bar chart are seen in Figs. 21 and 22.

Instead of the bar chart, comparisons are often made in terms of the areas of squares or circles, or of the volumes of cubes or spheres, as in Figs. 23 and 24. These devices, however, force the
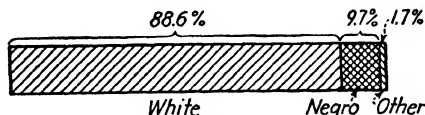
Fig. 21.—Percentage of the population of the United States represented by each race, 1930. (*Adapted from R. Clyde White, Social Statistics*, p. 178, *Harper & Brothers, New York*, 1933.)
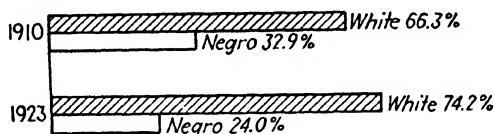
Fig. 22.—Percentage of white and Negro races among the commitments to prisons and reformatories, 1910 and 1923. (*From R. Clyde White, op. cit.*, p. 179.)
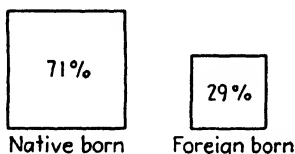
Fig. 23.—Ratio of native born to foreign born in City X: 1930.
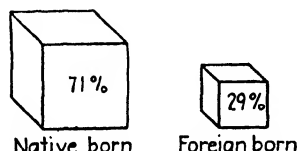
Fig. 24.—Ratio of native born to foreign born in City X: 1930.
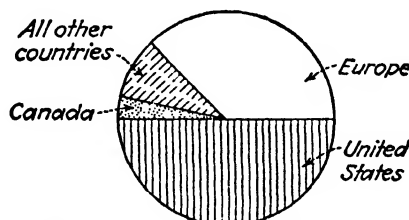
Fig. 25.—World distribution of telephones. (*Adapted from G. R. Davies and Dale Yoder, Business Statistics*, p. 40, *John Wiley & Sons, Inc., New York*, 1937.)

eye to perform the rather difficult feat of measuring two or even three dimensions simultaneously.

The so-called "pie chart," pictured in Fig. 25, is convenient for showing how a whole is subdivided.

MARCH 1929   MARCH 1931
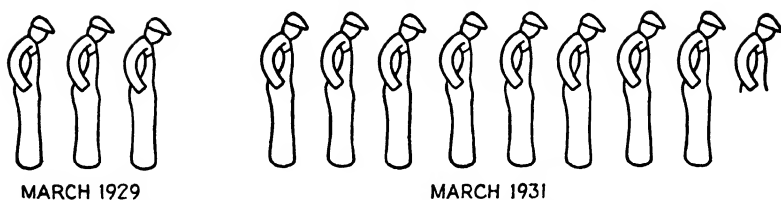
FIG. 26.—Estimated unemployment, United States, March, 1929, and March, 1931. (*Adapted from On Relief, Federal Emergency Relief Administration, Chart IX.*)

LEGEND
None
0.01%–20.0%
20.1%–40.0%
40.1% and over
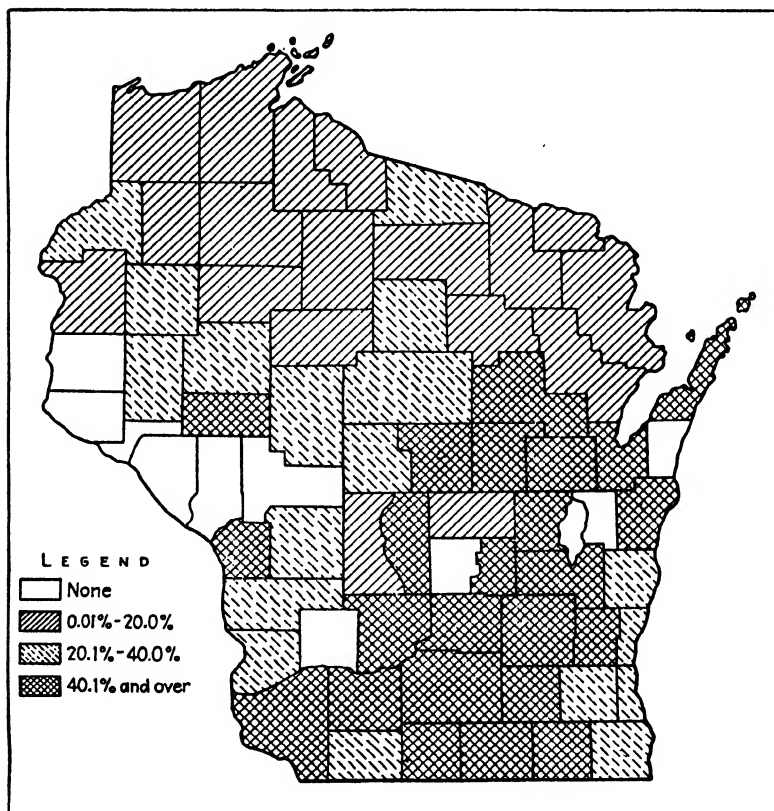
FIG. 27.—Percentage of unemployment relief expenditure paid locally, state of Wisconsin, year ending Sept. 1, 1934.

More realistic and striking than any of the preceding devices are pictograms, of which Fig. 26 is an example.

Maps are treated in many ingenious ways for statistical purposes. Crosshatching (see Fig. 27), the insertion of pictograms, and spotting are common devices.

In any attempt to present statistical figures in graphic form, the following two principles are to be kept in mind. (1) The graph should be more quickly and easily comprehended than the same data in tabular or nongraphic form. Graphs are sometimes so complex or ingenious that they can be deciphered only with the aid of the textual and tabular material that they are intended to clarify. (2) The graph should not misrepresent or exaggerate the facts.

Exercises

1. The following figures taken from the Fifteenth Census of the United States represent the growth in the population of Milwaukee:

1930	578,249	1880	115,587
1920	457,147	1870	71,440
1910	373,857	1860	45,246
1900	285,315	1850	20,061
1890	204,468	1840	1,712

Show these data graphically.

2. Suppose that you grade the behavior of a group of juvenile delinquents in a reform school and want to post a weekly chart showing the standing of each delinquent. Describe briefly the kind of chart you would use.

3. The charts on page 92 show how the numbers of the insane, epileptic, and feeble-minded persons in state institutions and the prison population in a certain state have increased in the last 25 years.

Have you any criticism of these charts?

4. Plot the distribution shown below as a frequency polygon. Show that the area under the polygon is equal to the total frequency, but that the area in some intervals is not equal to the frequency in those intervals.

DISTRIBUTION OF 106 EMPLOYEES BY AGE CLASS

Age of Employee, Years	Employees
15–24	14
25–34	49
35–44	23
45–54	13
55–64	6
65–74	1

1912	1927	1937
14,426	20,540	28,715

I N S A N E

| 1,479 | 4,562 | 6,995 |

FEEBLE MINDED AND EPILEPTIC

| 3,276 | 7,050 | 11,578 |

P E N A L
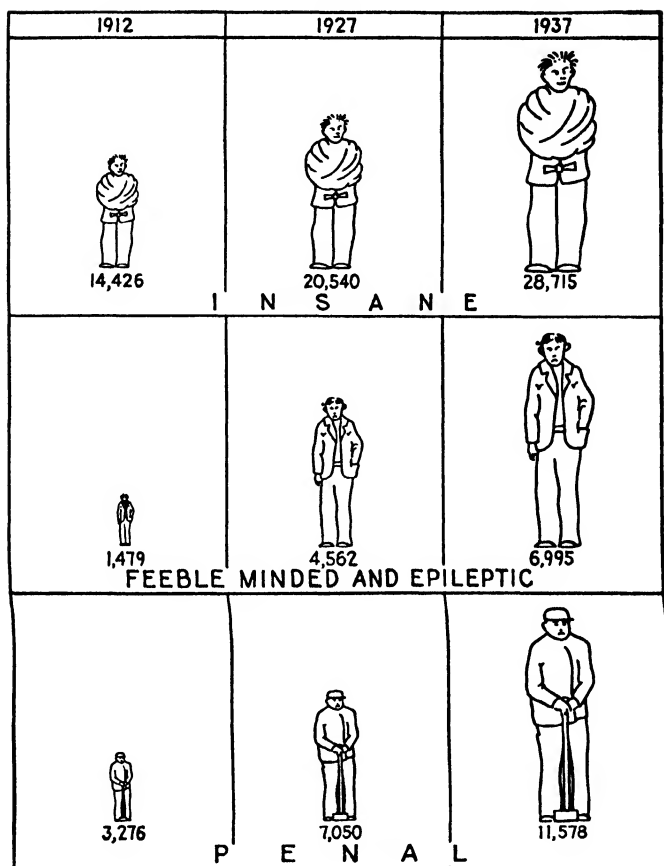
5. Rearrange the frequencies of question 4 in class intervals of 15–18, 19–22, 23–26, and so on.

6. Plot the data of Problem 4 as an ogive, and read off the age below which 75 per cent of the employees fall.

7. Devise a problem for which a Lorenz curve is suitable, graph the curve from your data, and show its use.

8. Plot the following data in such a way that (*a*) it gives an unbiased picture of the rate of change, (*b*) it exaggerates the impression of the rate of change.

POPULATION OF MADISON, WIS., 1890–1940*

Year	Population
1890	13,426
1900	19,164
1910	25,531
1920	38,378
1930	57,899
1940	66,802

* From the Fifteenth Census of the United States, Bureau of the Census.

References

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chaps. IV, V, VI, Prentice-Hall, Inc., New York, 1939.

KARSTEN, K. G.: *Charts and Graphs*, Prentice-Hall, Inc., New York, 1923.

MUDGETT, B. D.: *Statistical Tables and Graphs*, Part II, Houghton Mifflin Company, Boston, 1930.

CHAPTER VII

AVERAGES AND RATES

1. The Need for an Average.—An investigator is interested, let us say, in the height of the residents of a certain Swiss community in the United States, on the theory that they are taller than their relatives in the old country. Unable to measure the whole community of over 4,000 persons, he takes a random sample of, say, 182 adult males, and gets their heights as accurately as possible. He then finds himself with 182 individual measurements. What will he do with them? He may perhaps first arrange them in order of magnitude, to form an array. If no two of the measurements happen to be identical, he will still have 182 different measurements. In any case, it will be impossible for him to hold in mind all the separate values, and he will feel the need of some one figure by which to represent them. This need will be still greater when he attempts to determine whether or not the American group is taller than a similar group in the Old World, because some of the former will be taller than some of the latter, and vice versa. In his search for a single figure by which to represent the many, he will certainly arrive at the idea of calculating an average.

2. The Mode.—The simplest form of average is the *mode (Mo)*, which is merely the value in a series that occurs most often. If the heights are all different, there can be no mode in ungrouped data. If some persons are of the same height, however, a mode may occur in our array. We then choose as the mode the height that occurs the greatest number of times. For example, in the following array of the heights of 10 European Swiss males, 4 ft. 11 in., 5 ft. 3 in., 5 ft. 7 in., 5 ft. 8 in., 5 ft. 9 in., 5 ft. 9 in., 5 ft. 10 in., 5 ft. 11 in., 6 ft. 0 in., the mode is 5 ft. 9 in.; but of course the sample is too small to give much information about the modal height of European Swiss males in general. Whether or not a mode is convincing depends on how conspicuously the modal height stands out above the others in frequency of occurrence. If the height 5 ft. 7 in. occurs 10 times and the height 5 ft.

7.3 in. occurs nine times, it is not certain that one is significantly more frequent than the other.

The situation becomes clearer if we decide to óverlook slight differences in height, and combine our measurements in carefully chosen class intervals, as in Table 21. If the distribution is rather regular, we may by inspection then determine whether or not any one class interval has a sufficiently larger frequency than any other to be confidently regarded as the *modal class*. If so, that is usually all we need to know. In Table 21, col. (1), the modal interval is evidently 60 to 64 inches. In col. (2), the distribution has two modes, *i.e.*, it is bimodal, suggesting that it may contain both males and females. In such cases, it often helps to plot the data in the form of a frequency polygon or histogram, *e.g.*, Fig. 28.
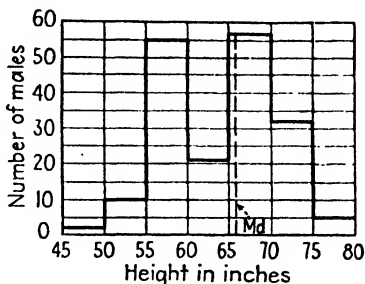
Fig. 28.—Histogram of bimodal frequency distribution of Table 21, col. (2).

TABLE 21.—HEIGHTS OF 165 AND 182 AMERICAN MALES OF SWISS DESCENT

Height, inches	Males	
	(1)	(2)
45–49	2	2
50–54	10	10
55–59	21	55
60–64	55	21
65–69	40	57
70–74	32	32
75–79	5	5
Total..........	165	182

Determination of the exact modal value in grouped data is complex, and cannot be treated here. Several rough methods of interpolating within the modal class are available, such as that of formula (1).

$$Mo = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) i. \tag{1}[1]$$

[1] Δ is the capital Greek letter *Delta*.

where L is the lower limit of the modal class, Δ_1 is the difference (disregarding signs) between the frequency of the modal class and the frequency of the class just below the modal class on the scale, Δ_2 is the difference (disregarding signs) between the frequency of the modal class and the frequency of the class just above the modal class, and i is the size of the modal class interval. Applying this formula to the distribution of Table 21, col. (1), we find the crude mode,

$$Mo = 60 + \frac{(55 - 21)(5)}{(55 - 21) + (55 - 40)}.$$
$$Mo = 63.5.$$

Another approximate method of finding the mode of a frequency distribution is provided by formula (2):

$$Mo = M - 3(M - Md). \tag{2}[1]$$

where M is the arithmetic mean of the distribution, and Md is the median, as described below. Assuming that for the distribution of Table 21, col. (1), $M = 64.68$, and $Md = 64.5$, formula (2) gives for the crude mode:

$$Mo = 64.68 - 3(64.68 - 64.5).$$
$$Mo = 64.1.$$

This value is a little different from that found by formula (1).

Mention of the conditions under which formula (2) is appropriate is made in the Sec. 5, below.

3. The Median.—Suppose that we have the following array of the heights of 11 American adult males of Swiss descent:

TABLE 22.—HEIGHTS OF 11 AMERICAN ADULT MALES OF SWISS DESCENT

Male	Height, inches
1	68
2	68.5
3	69
4	69.5
5	70
6	71
7	71.5
8	72
9	72.5
10	73
11	74

[1] See derivation of formula (54), Chap. IX.

A quick way of getting some idea of an average for this series would be to note the height that stands at the middle of the series. This is seen to be 71 inches, or height number 6 in rank order. This kind of average is called the *median*, which is defined as the middle value, or that value which is exceeded by as many values as it exceeds.

Now if a twelfth person of, say, height 75 inches is added to the above group, a difficulty arises. There is no middle value. Unless we are willing to take the mean height of the sixth and seventh persons $\left(\dfrac{71 + 71.5}{2} = 71.25\right)$ as the median, we must say that there is none. Although the median so found, 71.25 inches, is a height that does not actually appear in the series, it is customary for most purposes to accept it as the median.

Consider another common case. Let the height of the fifth person in the first group of 11 persons be 71 inches. Again, strictly speaking, there can be no median that meets the definition, because there are no longer as many heights below the middle height as there are above it. As before, a compromise is commonly made by taking the middle value (71 inches) as the median.

From the above, it will be noticed that the formula for locating the median value in an *ungrouped* series is to add one to the number of values and divide by two:

$$\frac{N + 1}{2}. \tag{3}$$

Thus, above, where $N = 11$, the position of the median value is $\dfrac{11 + 1}{2} = \dfrac{12}{2} = 6$, or the value in position 6; and where $N = 12$, $\dfrac{12 + 1}{2} = \dfrac{13}{2} = 6.5$, the median value is the height in position 6.5, which can only be the mean of the heights in positions 6 and 7.

The above relates to ungrouped data. When the items of a series are *grouped* in class intervals, the median is regarded as the value on the X scale that divides the area of the frequency histogram or curve into two equal parts, as shown in Fig. 28. Thus, in Table 21, col. (2), $N/2 = 182/2 = 91$. Now, 88 frequencies fall below the class limit, 65 inches, so that $91 - 88 = 3$ frequencies fall inside the interval 65 to 69. Since there are 57

frequencies in this interval, and the width of the interval is 5 inches, the median falls $\frac{3}{57} \times 5 = 0.263$ inch inside the interval, or at the point $65 + .263 = 65.263$ inches on the X scale. The area below the median is then $2 + 10 + 55 + 21 + \frac{.263}{5.000}$ $(57) = 91$, that above the median is $\frac{4.737}{5.00}(57) + 32 + 5 = 91$, and the two areas are equal.

The simplest way to find the median of grouped data is as follows: Accumulate the frequencies, as in the last column of Table 23. Divide N by 2: $165/2 = 82.5$. Look down the column of accumulated frequencies until the frequency in the position 82.5 is found, in the interval 60–64. From 82.5 subtract the accumulated number of frequencies below the median interval: $82.5 - 33 = 49.5$. Multiply the width of the class interval by the fraction 49.5/55, formed by the difference just found as numerator and the frequency of the median interval as the denominator: $5 \times 49.5/55 = 4.5$. Add this quotient to the lower limit of the median interval: $60 + 4.5 = 64.5$. This is the median height for the table.

TABLE 23.—HEIGHT OF 165 AMERICAN ADULT MALES OF SWISS DESCENT

Height, inches	Males	
	Number	Accumulated number
45–49	2	2
50–54	10	12
55–59	21	33
60–64	55	88
65–69	40	128
70–74	32	160
75–79	5	165
Total..........	165	

We can express the above steps by means of a formula, which is applicable to frequency distributions:

$$Md = L + \left(\frac{\frac{N}{2} - F}{f}\right)i, \tag{4}$$

where L is the lower limit of the class interval in which the median falls, F is the number of accumulated frequencies that fall below (*i.e.*, in class intervals with limits smaller than those of) the median class interval, f is the number of frequencies in the median class interval, i is the size of the median class interval, and N is the total frequency of table. $N/2$ is first found, and then the remaining symbols can be evaluated and substituted in the formula, as indicated in the preceding paragraph. Thus, for the problem above, $Md = 60 + \left(\dfrac{\frac{165}{2} - 33}{55}\right) 5 = 64.5$, as before.

4. The Arithmetic Mean.—The arithmetic mean, M, is the type of average that is most often used. It is the sum of the X values divided by their number, N:

$$M = \frac{\Sigma X}{N}. \tag{5}*$$

For example, in the case of the ungrouped values, 3, 7, 2, 12, 1, 16, 4, representing the numbers of children in seven Italian immigrant families, their sum is 45, and there are seven of them, so that $M = 45/7 = 6.43$.

If some of the above values had occurred more than once, we might have

X	X (*Continued*)
1	4
2	7
2	7
3	7
3	12
3	12
4	16
4	$\underline{16}$
4	$\overline{111}$
4	

$$M = \frac{111}{18} = 6.17$$

But, as shown in Chap. V, this long array may be condensed:

* The Greek letter, Σ, capital Sigma, means to *sum*, or add, the X values.

TABLE 24.—NUMBER OF CHILDREN IN 18 ITALIAN IMMIGRANT FAMILIES

Children (X)	Families (f)*	fX†
1	1	1
2	2	4
3	3	9
4	5	20
7	3	21
12	2	24
16	2	32
Total	18	111

* Frequency.
† Frequency multiplied by X.

In the case of grouped data, it is more convenient to write formula (5) in the form

$$M = \frac{\Sigma fX}{N},\tag{6}$$

where f is the frequency.

Substituting in formula (6) $N = 18$ and the total of the third column of the array just above,

$$M = \tfrac{111}{18} = 6.17,$$

as before.

Formula (6) may be applied to any frequency distribution, *e.g.*, that of Table 25.

TABLE 25.—HEIGHT OF 165 AMERICAN ADULT MALES OF SWISS DESCENT

Height		Adult males	
Inches	X*	f	fX
45–49	47.5	2	95.0
50–54	52.5	10	525.0
55–59	57.5	21	1207.5
60–64	62.5	55	3437.5
65–69	67.5	40	2700.0
70–74	72.5	32	2320.0
75–79	77.5	5	387.5
Total...........		165	10,672.5

* Mid-points.

$$M = \frac{10,672.5}{165} = 64.68.$$

As pointed out earlier, the mean calculated from a frequency table in which the mid-points are not identical with the means within the intervals is, of course, somewhat inaccurate, as is any other average or statistic found from such a table.

It is possible to simplify the calculations needed to find the arithmetic mean (usually called simply the mean) in a frequency distribution such as that of Table 25. Suppose that the mid-points of any distribution are X_1, X_2, X_3, etc. They can be
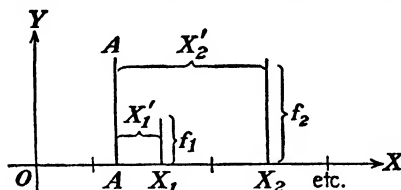
Fig. 29.—Diagram used in derivation of formula (13).

represented by the above diagram, where f_1 is the frequency in the X_1 interval, etc., measured along the Y axis.

By formula (6),

$$M = \frac{\Sigma f X}{N}$$

But suppose that we choose to measure the X values from some arbitrarily assumed or "guessed" point on the X axis, say A, in Fig. 29. Then

$$X_1 = A + X_1', \tag{7}$$
$$X_2 = A + X_2',$$

where the X''s represent the distances of the X's measured from A.

If we further choose to reduce the size of the X''s by dividing them by the size of the class interval, i, or other constant, we have

$$\frac{X_1'}{i} = d_1, \tag{8}$$

$$\frac{X_2'}{i} = d_2, \text{ etc.,}$$

or

$$X_1' = d_1 i, \tag{9}$$
$$X_2' = d_2 i, \text{ etc.}$$

Substituting the values of X_1' and X_2' from (9) in (7),

$$X_1 = A + d_1 i, \tag{10}$$
$$X_2 = A + d_2 i, \text{ etc.}$$

Substituting from (10) in (6),

$$M = \frac{\Sigma f(A + di)}{N}, \tag{11}$$

$$M = \frac{\Sigma fA}{N} + \frac{\Sigma fdi}{N}.$$

Constants can always be placed outside the summation sign,[1] so that

$$M = \frac{A \Sigma f}{N} + \frac{i \Sigma fd}{N}.$$

Now

$$\Sigma f = N.$$

So that

$$M = A \frac{N}{N} + \frac{i \Sigma fd}{N},$$

or

$$M = A + \frac{i \Sigma fd}{N}. \tag{13}$$

Let us apply formula (13) to find the mean of the frequency distribution of Table 26:

TABLE 26.—HEIGHT OF 165 AMERICAN ADULT MALES OF SWISS DESCENT

X*	$X' = X - A$	$d = \frac{X'}{i}$	f	fd
47.5	−15	−3	2	− 6
52.5	−10	−2	10	−20
57.5	− 5	−1	21	−21
62.5	0	0	55	0
67.5	+ 5	+1	40	+40
72.5	+10	+2	32	+64
77.5	+15	+3	5	+15
Total........			165	+72

* Mid-points.

In the above table, by arbitrary choice, the assumed mean is

$$A = 62.5.$$
$$i = 5.$$
$$\Sigma fd = 72.$$
$$N = 165.$$

[1] Notice the principle that

$$\Sigma(X + Y) = \Sigma X + \Sigma Y \tag{12}$$

Therefore, substituting in formula (13),

$$M = 62.5 + \frac{5(72)}{(165)},$$
$$M = 62.5 + 5(.436),$$
$$M = 62.5 + 2.18,$$
$$M = 64.68,$$

which is the same as we found the mean to be by the "long" method. The calculations required in Table 26 are greatly reduced compared with those in Table 25.

The second column of Table 26 is inserted for explanatory purposes only, and is omitted except when irregular class intervals cause difficulties. Table 27 illustrates the usual form for computation.

TABLE 27.—NUMBER OF RELIEF CASES PER BLOCK IN A SLUM AREA OF A CITY

X	d	f	fd
1.5	−3	4	−12
3.5	−2	10	−20
5.5	−1	14	−14
7.5	0	26	0
9.5	+1	19	+19
11.5	+2	14	+28
13.5	+3	8	+24
15.5	+4	4	+16
17.5	+5	1	+ 5
Total............		100	+46

$$M = 7.5 + 2\left(\frac{46}{100}\right) = 8.42$$

Notice that it makes no difference in the result where the $d = 0$—*i.e.*, the assumed mean—is placed. A good way to check the work is to perform the calculations from two different assumed means.

Formula (13) also holds for irregular class intervals if i, which may be any convenient divisor as well as the size of the class interval, is held constant. This may be illustrated below.

Consider this table of age distribution for Wisconsin, from the 1930 census.

TABLE 28.—AGE DISTRIBUTION OF POPULATION, WISCONSIN, 1930

Age	Per cent f	X^*	$X - A$	$\dfrac{X - A}{5} = d$	fd	Accumulated f†
Under 5.........	9.2	2.5	-20.0	-4.0	-36.8	9.2
5–9.............	9.9	7.5	-15.0	-3.0	-29.7	19.1
10–14...........	9.7	12.5	-10.0	-2.0	-19.4	28.8
15–19...........	9.2	17.5	$-\ 5.0$	-1.0	$-\ 9.2$	38.0
20–24...........	8.3	22.5	0	0	0	46.3
25–29...........	7.7	27.5	$+\ 5.0$	$+1.0$	$+\ 7.7$	54.0
30–34...........	7.4	32.5	$+10.0$	$+2.0$	$+14.8$	61.4
35–44...........	14.0	40.0	$+17.5$	$+3.5$	$+49.0$	75.4
45–54...........	10.6	50.0	$+27.5$	$+5.5$	$+58.3$	86.0
55–64...........	7.2	60.0	$+37.5$	$+7.5$	$+54.0$	93.2
65–74...........	4.6	70.0	$+47.5$	$+9.5$	$+43.7$	97.8
75 and over.....	2.0	?	?	?	?	99.8
Total.........	99.8	....	......	.....	132.4	

* Mid-points. The census records age in whole years, as of the last birthday. But since the actual ages are not discrete, age should be treated as continuous. Otherwise, all averages will be too low.

† Accumulated frequencies.

Finding the mean for the table below age 75[1] by substituting in formula (13),

$$M = 22.5 + 5\left(\frac{132.4}{97.8}\right) = 22.5 + 5(1.354)$$
$$M = 29.27$$

In such a table as this, with an open interval, only the median and mode can be found for the total table. Why? What is the median value for Table 28?

5. Interpretation of the Common Averages.—The arithmetic mean, M, is the most familiar type of average. It is amenable to algebraic operations which cannot be applied to the median or mode. Suppose we know that the mean of one distribution of 50 items is 4, and the mean of a second comparable distribution of 75 items is 6. Then the mean of both distributions is $\dfrac{(4 \times 50) + (6 \times 75)}{50 + 75} = 5.2$. The only accurate way of finding

[1] The mean, of course, cannot be found for the table including the open interval, "75 and over," because no mid-point can be assigned to an open interval.

the median of the total distribution is actually to combine the distributions, interval by interval, and recompute the median and mode for the combined distribution, just as was done for the separate distributions. If there are several medians given, it is possible to find the median median, but it is not likely to be the same as the median of the combined distributions. Although the mean of two or more medians is sometimes used, the meaning of such a combination of averages is not clear. "A correct total cannot be obtained by multiplying the median by the number of items" in a distribution.[1]

A second characteristic of the mean is that it alone of the three averages reflects the exact value of every item. If extreme values occur in a series, they affect the mean much more than the median or the mode, because the median is affected only by the circumstance that an item is greater or smaller than the median item—the amount of the difference being of no consequence—and the mode is affected only by whether or not the size of a value throws it into one class interval or another. Consider the series of ages in years, 2, 4, 7, 10, 13, 15, 19. $M = 10$, $Md = 10$. If the three items that are larger than 10 are replaced by three others also larger than 10, the Md stays the same, but the M changes. Thus for 2, 4, 7, 10, 58, 70, 80, $M = 33$, $Md = 10$. This is sometimes an advantage of the mean, and sometimes a disadvantage. If the extreme values are regarded as atypical of the series, the median will be a better average than the mean, because the median is less influenced by such values. If, on the other hand, the extreme values are thought to be an integral part of the series and to deserve full weight, then the mean is more appropriate than the median. In series where the mean seems inappropriate, it is often advisable to question the representativeness of any average, and to drop the atypical items.

A third important trait of the mean is that it usually changes less than the other two averages, from sample to sample. Suppose that the I.Q.'s of the first 100 students met on a college campus are taken and the M, Md, and Mo of these I.Q.'s are computed. The same thing is done with a second hundred students, a third, and so on. Then the differences between the means of the several samples will generally be less than the differences

[1] WILFORD I. KING, *The Elements of Statistical Method*, p. 131, The Macmillan Company, New York, 1918.

between the medians or the modes.   This sampling stability of the mean is very much in its favor.

For such reasons as the above, in the averaging of measurements the arithmetic mean is always to be preferred to the median or mode unless it is felt to be much less representative of the series than they are, or unless, because of open-end class intervals, the mean cannot be calculated.   When we have to deal with a series of ranked items, rather than measured values, however, only the median applies.

A frequency distribution is exactly balanced along the perpendicular erected at the mean.   The sum of the deviations of a series of values from their mean with regard for signs, *i.e.*, the algebraic sum, is always zero.   This is not true of the other averages except in perfectly symmetrical distributions, where the mean, median, and mode all coincide (see Fig. 30).   A distribution is symmetrical when equal frequencies occur at equal distances above and below the mean, as in Table 29 and Fig. 30.

Fig. 30.—Graph of symmetrical frequency distribution of Table 29.

TABLE 29.—DISTRIBUTION OF EXPECTED NUMBER OF HEADS FROM 1,024 TOSSES OF 10 COINS EACH

Heads among 10 Coins	Tosses of 10 Coins
0	1
1	10
2	45
3	120
4	210
5	252
6	210
7	120
8	45
9	10
10	1
Total................	1,024

On the other hand, when signs are disregarded, the sum of the deviations is least in the case of the median.

Typically, in distributions that are not symmetrical or bell-shaped, but skewed, *i.e.*, extending farther on one side than on the other, the mean is pulled farthest in the direction of the skewness (because of its sensitiveness to extreme values), the mode is nearest the end of the scale opposite the direction of the skewness, and the median falls somewhere in between the other two (see Fig. 31). Indeed, in moderately skewed distributions, the median is generally about one-third of the distance from the mean to the mode, a fact utilized in formula (2) above. If the three averages are calculated for the skew distribution of Table 28 below age 75, using formula (2) for the mode, they will be found to fall in this way ($M = 29.27$; $Md = 27.34$; $Mo = 23.48$).

Fig. 31.—Skewed frequency distributions.

The usefulness of any average usually depends upon how representative it is of its distribution or series, *i.e.*, upon what proportion of the items in the series is close to the average. Although it is mathematically possible to calculate the mean, median, or mode for any series, the concept of the average as a value representative of the series has much more validity in the case of some series than of others. It is most valid for symmetrical distributions, and least valid for distributions shaped like the letter J (or reversed J), or the letter U, illustrated in Table 30, cols. (2) and (3), respectively, and Figs. 33, 34. In the case of J and U shaped distributions, any average is likely to conceal more important information than it reveals, and for this reason it is usually advisable not to compute averages for distributions of

TABLE 30.—AGE DISTRIBUTIONS (HYPOTHETICAL DATA)

Years of age	(1) f	(2) f	(3) f
0– 4.9	21	18	116
5– 9.9	53	21	53
10–14.9	116	47	18
15–19.9	47	53	47
20–24.9	18	116	132

such extreme types. Perhaps the mode is the best of the three averages in situations of this kind; but even its value is questionable.

Enlarging on the last point, special precaution is necessary to avoid the use of averages to represent a group that varies widely within itself. Thus a single infant mortality rate for a county containing a large city and a rural area in which the rates are very different is likely to be not only meaningless, but misleading. This point must be kept constantly in mind in most statistical problems, *e.g.*, the calculation of a correlation coefficient. The latter, which is an average, may indicate a moderate amount of
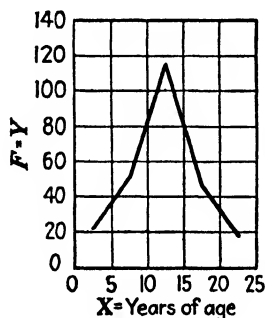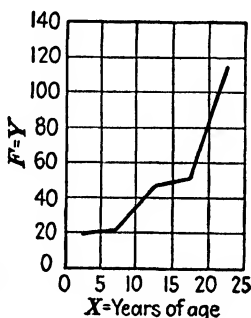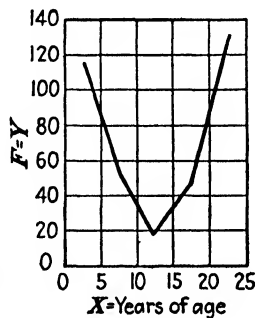
Fig. 32.   Fig. 33.   Fig. 34.

Fig. 32.—Graph of roughly symmetrical frequency distribution of Table 30, Col. (1).

Fig. 33.—Graph of J-shaped frequency distribution of Table 30, Col. (2).

Fig. 34.—Graph of U-shaped frequency distribution of Table 30, Col. (3).

relationship over the whole table, whereas actually there is no relationship at one end of the table and a close relationship at the other (see Chap. X).

It should be noticed that an average, usually the mean, may sometimes legitimately be used for the purpose of resolving a series of values into a single composite value, whether the latter is "representative" of the values in the series or not. This is the case when the chief interest lies merely in comparing the composite values of two or more series, as the mean size of income of all workers with the mean size of income of unskilled laborers alone.

In most cases, it is important to exhibit the table of the frequency distribution as a whole, so that the distribution of the items, as well as their averages, may be known to the reader.

It is also a practice in doubtful cases to present all three averages side by side, so that their differences may be seen. This, however, may merely throw upon the reader the responsibility of choosing an average.

6. The Geometric Mean.—In averaging a series of numbers that bear an approximately constant ratio to one another, like 2, 4, 8, 16, none of the three averages described above is as appropriate as the *geometric mean*. The geometric mean is used to average any series in which changes are expressed as rates rather than as absolute differences. It is also preferable for averaging some skewed distributions, since it gives less weight to extreme variations than does the arithmetic mean.

The geometric mean is always smaller than the corresponding arithmetic mean. When a series contains a zero or negative value, its geometric mean cannot be found. Just as the sum of the plus deviations is equal to the sum of the minus deviations from the arithmetic mean, so the product of the ratios of the values smaller than the geometric mean to the geometric mean is equal to the product of the ratios of the geometric mean to the values larger than the geometric mean (*e.g.*, the geometric mean of 5, 8, 10, and 12 is 8.3, and $\frac{5}{8.3} \times \frac{8}{8.3} = \frac{8.3}{10} \times \frac{8.3}{12}$). Also, corresponding to the fact that when each member of a series is replaced by the arithmetic mean of the series the sum of the series is not changed (*e.g.*, $3 + 7 + 5 = 15$, and $5 + 5 + 5 = 15$), so, when each member of a series is replaced by the geometric mean, the product remains the same (*e.g.*, $12 \times 34 \times 4 = 1,632$, and $11.7735 \times 11.7735 \times 11.7735 = 1,632$).

For an ungrouped series of values $X_1, X_2, \ldots X_n$, the formula for the geometric mean is

$$G = \sqrt[n]{X_1 \cdot X_2 \cdots X_n}. \tag{14}$$

For grouped data,

$$G = \sqrt[n]{X_1{}^{f_1} \cdot X_2{}^{f_2} \cdots X_n{}^{f_n}}, \tag{15}$$

where X_1 is a mid-point and f_i, its exponent,[1] is the corresponding class frequency. Computation, however, is most conveniently

---

[1] *Exponent* means the power to which X is raised, *e.g.*, $(X)^2$. Here the exponent is 2, the second power.

done by means of logarithms, using the respective formulas:

$$\log G = \frac{1}{n} \sum_{1}^{n} \log X_i \qquad (16)$$

$$\log G = \frac{1}{N} \sum_{1}^{n} f_i \log X_i, \qquad (17)$$

where $N = \sum_{1}^{n} f_i$.

To illustrate the use of formula (16), the geometric mean of the rates in col. (4) of Table 32 below is found from a table of logarithms:[1]

$$\begin{aligned}
\log G &= \tfrac{1}{7}(\log 0.015 + \log 0.058 + \log 0.061 + \log 0.047 \\
&\qquad\qquad + \log 0.029 + \log 0.011 + \log 0.001) \\
&= \tfrac{1}{7}(8.17609 - 10 + 8.76343 - 10 + 8.78533 - 10 \\
&\qquad + 8.67210 - 10 + 8.46240 - 10 + 8.04139 \\
&\qquad\qquad\qquad\qquad\qquad - 10 + 7.00000 - 10) \\
&= \tfrac{1}{7}(57.90074 - 70) \\
&= 8.27153 - 10 \\
G &= 0.019
\end{aligned}$$

Notice that the geometric mean obtained by formula (16) is unweighted, *i.e.*, each rate is given equal weight. The unweighted arithmetic mean of the same rates is 0.03171, while the weighted arithmetic mean rate, from cols. (2) and (3) of Table 32, is 26,326/790,193 = 0.03331.

The total column of the table in Exercise 1 below shows a skewed distribution, so that the geometric mean should be more representative of it than the arithmetic mean. By formula (17),

$$\begin{aligned}
\log G &= \tfrac{1}{391}(73 \log 2 + 96 \log 6 + 101 \log 10 + 48 \log 14 \\
&\qquad\qquad\qquad\qquad + 52 \log 18 + 21 \log 22) \\
&= \tfrac{1}{391}[73(0.30103) + 96(0.77815) + 101(1) + 48(1.14613) \\
&\qquad\qquad\qquad\qquad + 52(1.25527) + 21(1.34242)] \\
&= \tfrac{1}{391}(21.97519 + 74.70240 + 101.00000 + 55.01424 \\
&\qquad\qquad\qquad + 65.27404 + 28.19082) = 0.88531 \\
G &= 7.68
\end{aligned}$$

The arithmetic mean is 9.72 and the median is 9.05. Formula (17) gives a weighted geometric mean, or geometric mean of a frequency distribution.

[1] See Appendix, Table 7, and accompanying explanation.

Notice the application of the geometric mean to the problem of estimating the population midway between two decennial censuses. In Table 31 the population of the United States in millions is shown at 10-year intervals from 1790 to 1940. When

TABLE 31.—POPULATION OF THE UNITED STATES, 1790–1940
(In millions)

Year	Population	Year	Population
1790	3.93	1870	38.56
1800	5.31	1880	50.16
1810	7.24	1890	62.95
1820	9.64	1900	75.99
1830	12.87	1910	91.97
1840	17.09	1920	105.71
1850	23.19	1930	122.78
1860	31.44	1940	131.41 (prelim.)

these figures are plotted, we get the absolute growth curve shown in Fig. 35. Now suppose it is wanted to estimate the
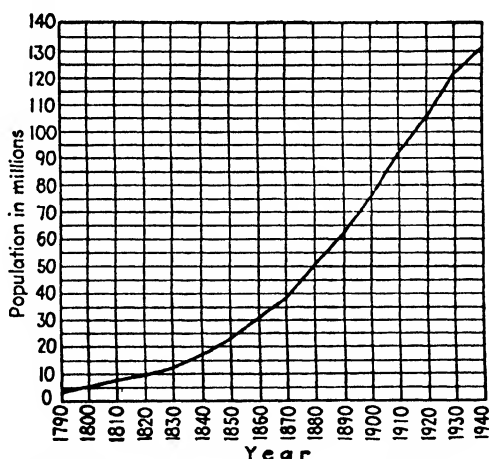
FIG. 35.—Absolute growth of population, United States, 1790–1940.

population in 1795, midway between the censuses of 1790 and 1800. If we take the arithmetic mean of the populations at 1790 and at 1800, we have

$$\frac{5.31 + 3.93}{2} = 4.62 \text{ millions.}$$

This evidently assumes that the absolute amount of population increase is the same over equal periods of time, since

$$4.62 - 3.93 = 0.69,$$

and $5.31 - 4.62 = 0.69$. From Table 31, however, we see that the differences, $5.31 - 3.93 = 1.38$ and $7.24 - 5.31 = 1.93$, are not equal, and this is borne out by inspection of Fig. 35. Actually, around the dates 1790 and 1800, as far as we can judge from the given data, the absolute growth in population was increasing. Under these conditions, the growth curve between 1790 and 1800 would probably be concave, as shown by line *a* in Fig. 36. The population in 1795 would then be somewhat less than that found by the method of the arithmetic mean,
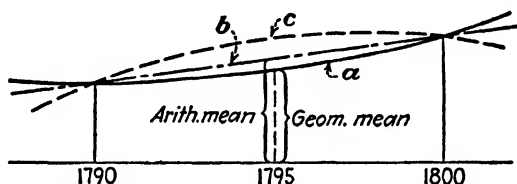
Fig. 36.—Probable trend of population growth in the United States, 1790–1800.

which implies a straight line rather than a concave trend (line *b* in Fig. 36). On the simple assumption that the rate of annual increase was constant between 1790 and 1800, the growth curve will be concave, and the geometric mean will give the exact population in 1795. The geometric mean is, therefore, usually regarded as the logical average to use when the growth curve is concave. The formula may be written

$$\bar{P} = \sqrt{P_0 P_{10}} = (P_0 P_{10})^{\frac{1}{2}}, \tag{18}$$

where $\bar{P}$ is the population midway between the two censuses, P_0 is the population at the first census, and P_{10} is the population at the second census. Substituting in this formula,

$$\bar{P} = \sqrt{3.93(5.31)} = 4.57 \text{ millions.}$$

The geometric mean is not a suitable average, however, when the absolute amount of change is less each decade, as happened between 1930 and 1940. The growth curve is then convex (like *c*, Fig. 36), so that both the arithmetic and the geometric means give too low estimates of the population midway between censuses.

If we wish to calculate the constant annual rate of population increase that was assumed in finding the geometric mean ($= 4.57$) above, we apply the formula

$$r = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}} - 1. \tag{19}$$

If, as before, $P_0 = 3.93$, $P_n = 5.31$, and $n = 10$, we have

$$r = \left(\frac{5.31}{3.93}\right)^{\frac{1}{10}} - 1 = (1.35)^{\frac{1}{10}} - 1.$$

By logarithms,

$$\log (1.35)^{\frac{1}{10}} = \tfrac{1}{10} \log 1.35 = \tfrac{1}{10}(0.13033) = 0.013033.$$

So

$$(1.35)^{\frac{1}{10}} = 1.03,$$

and $r = 1.03 - 1.00 = 0.03$.

That is, in finding the geometric mean we assumed that the population increased at the average rate of about 3 per cent per year between 1790 and 1800.

For the same problem, the arithmetic mean gives a rate of $\dfrac{5.31 - 3.93}{3.93(10)} = 3.5$ per cent, which if assumed to be constant over the 10-year period would result in a population in 1800 of

$$P_{10} = P_0(1 + r)^{10}. \tag{20}{[1]}$$
$$P_{10} = 3.93(1.035)^{10}.$$

$$\log (1.035)^{10} = 10 \log 1.035 = 10(0.01494) = 0.14940.$$

So

$$(1.035)^{10} = 1.411,$$

and

$$P_{10} = 3.93(1.411),$$

or

$$P_{10} = 5.55 \text{ millions},$$

whereas, actually, $P_{10} = 5.31$ millions.

[1] Formulas (19) and (20) may be derived as follows:
Let

P_0 = Population of the state on Jan. 1, 1930,
P_1 = Population of the state on Jan. 1, 1931; etc.
r = constant annual rate of increase.

7. Population Rates.—The ratio of divorces to population, say 3.2 per 1,000, is an illustration of the kind of *rate* that is important for sociologists. Other examples are the crude birth rate (births per 1,000 population per year) and the crime rate (say, convictions per 1,000 males 10 years old and over per year). A rate shows the amount of one variable per given amount of another variable *e.g.*, the number of births in relation to a given number of women of child-bearing age in a population.

In working with population rates, such as marriage rates, death rates, etc., it is helpful to have in mind what is meant by a rate. Mathematicians define a rate as the amount of change in a function (dependent variable) that occurs per unit change in the independent variable. The rate of travel of an automobile is the number of miles by which its position in space

Then

$$P_1 = P_0 + P_0 r = P_0(1 + r),$$
$$P_2 = P_1 + P_1 r$$
$$= P_0(1 + r) + P_0(1 + r)r$$
$$= P_0(1 + r)(1 + r)$$
$$= P_0(1 + r)^2,$$

$$P_3 = P_2 + P_2 r$$
$$= P_0(1 + r)^2 + P_0(1 + r)^2 r$$
$$= P_0(1 + r)^2(1 + r)$$
$$= P_0(1 + r)^3.$$

Similarly,

$$P_{10} = P_0(1 + r)^{10}$$

If n = number of years between censuses,

$$P_n = P_0(1 + r)^n,$$

or

$$(1 + r)^n = \frac{P_n}{P_0},$$

$$\log (1 + r)^n = \log \frac{P_n}{P_0},$$

$$n \log (1 + r) = \log \frac{P_n}{P_0},$$

$$\log (1 + r) = \frac{1}{n} \log \frac{P_n}{P_0},$$

$$\log (1 + r) = \log \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}},$$

or

$$1 + r = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}},$$

and

$$r = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}} - 1.$$

(function) changes per change of 1 hour in time (independent variable). How does a marriage rate fit the mathematical idea of a rate? The usual form of the marriage rate is the number of marriages per 1,000 population per year. Here, however, there are three variables instead of the two mentioned in the mathematical definition of a rate. Which of these is the function, which is the independent variable, and how is the third variable to be interpreted? The time element is usually regarded as the independent variable in rate problems, and the factor that varies with time, as the function. In our example, both the number of marriages and the size of the population base may change from year to year. Either of these alone related to time would give a mathematical rate. But we are not interested in such a rate. Rather, we want to know how the ratio of marriages to total population changes with time. It is, then, this *ratio* that is the function in our marriage rate.

In the case of the marriage rate, we are primarily interested in the annual changes in the number of marriages, and not in the change in the population base. The only reason for introducing the population base at all is to eliminate it as a cause of change in the number of marriages, so that the annual change in the number of marriages may be comparable from one population to another.

This raises an important point. Is the population base the only factor that needs to be eliminated or controlled in order that the marriage rate may mean just what we want it to? In order to have the marriage rate as comparable as possible from one population to another, should we not also control the factors of age and sex composition, so that their influences are removed from the rate? That depends on the question we want to answer. If our question is, which of two or more total populations has the higher marriage ratio, regardless of the causes involved, we do not control age and sex in our ratio. But if we wish to know which of the populations would have the higher marriage rate *if* their age and sex distributions were the same, we must control age and sex. This leads us to the so-called age-specific, gross, and net marriage rates for females. In all such rates, we note as a general principle that the denominator or base of the final rate should ideally contain only the group *exposed* to the event (*e.g.*, if the event is marriage, the group

exposed should be composed exclusively of, say, *unmarried* females), while the numerator should contain the number of *events* (*e.g.*, marriages) occurring in the year.   In the case of most crude rates, like official birth and marriage rates, this principle is disregarded.

When marriage or other rates are plotted, it is usually advisable to plot them on semilogarithmic paper, in order that the rate of change may be shown by the *steepness* of the graph.   Plotting the rates directly on semilogarithmic paper is equivalent to plotting the logarithms of the rates, which in turn is similar to plotting the *percentage of change* in the rates from year to year (see Chap. VI, Fig. 19).

It may be of interest to compute two of the most important of the refined rates used in current vital statistics.   The *gross reproduction* rate is defined as the average number of girls born per woman passing through child-bearing age, say 15 to 50 years, without mortality, and exposed to the birth rate of a given year. The *net reproduction rate* is simply the gross reproduction rate corrected for mortality.   In Table 32 these rates have been found for Wisconsin, with the year 1934 as the base.   The gross rate appears as the total of col. (5), and the net rate as the total of col. (7).   It is seen that each 1,000 women born, if none died and all were subjected to the average age-specific rates of 1934, would bear 1,110 daughters.   However, if these 1,000 women were exposed to the death rates found in an appropriate life table, they would bear only 995 daughters to start the next generation. Since the actual distribution of women by age groups was eliminated as a factor in Table 32 from col. (4) on, it is possible that there may be a disproportionate number of young females in the population of Wisconsin in 1934 and that this may prevent the population from actually declining for a time, even though the net reproduction rate is less than 1.   But if the net reproduction rate of 1934 should continue until the age distribution was stabilized, the female population of the state would then begin to decrease at the rate of 5 per 1,000 per generation.   As a matter of fact, the birth rate was unusually low in 1934 on account of the economic depression and has since risen somewhat.   The average birth rate over a period of, say, 3 to 5 years furnishes a more stable base than the rate for a single year, and for some purposes should be preferred.

TABLE 32.—GROSS AND NET REPRODUCTION RATES IN WISCONSIN, 1934

Age groups	Females, 15–49, July 1, 1934	Female live births, 1934	Daughters born per female, 15–49, 1934	Average daughters born to female, 15–49, in 5-year period	Female survival rates from birth	Average daughters born to a female surviving to age 50
(1)	(2)*	(3)†	(4)‡	(5)§	(6)‖	(7)¶
15–19	139,600	2,147	0.015	0.075	0.92512	0.070
20–24	131,369	7,599	0.058	0.290	0.91480	0.265
25–29	118,042	7,256	0.061	0.305	0.90117	0.275
30–34	108,496	5,090	0.047	0.235	0.88626	0.208
35–39	103,165	2,987	0.029	0.145	0.87016	0.126
40–44	101,089	1,147	0.011	0.055	0.85071	0.047
45–49	88,432	100	0.001	0.005	0.82522	0.004
Total......	790,193	26,326		1.110		0.995

* Estimated from the 1930 census with the aid of a life table for Wisconsin.

† Found by applying the percentage of total births, female, in 1934 to total live births, corrected for underregistration.

‡ Column (3) divided by col. (2).

§ Column (4) multiplied by 5, since a woman in any 5-year age group is assumed to bear as many daughters in each of the 5 years as in 1934.

‖ Taken from Life Table for White Females in Wisconsin, 1929–1931, prepared by the Metropolitan Life Insurance Company.

¶ Column (5) multiplied by col. (6).

Exercises

NOTE: A calculating machine will save time in solving the problems in this text.   At least the student should own an inexpensive slide rule.

1. *a.* Find the crude mode, where appropriate, of each of the following four series, and of all four combined, using formula (1):

AGE OF CHILDREN IN FOUR THREE-GENERATION KINSHIP GROUPS

Age of child, years	Number of children in kinship group				Total children
	I	II	III	IV	
(X)*	(f₁)†	(f₂) ·	(f₃)	(f₄)	(f)
2	10	15	30	18	73
6	21	20	29	26	96
10	33	36	18	14	101
14	20	10	6	12	48
18	12	7	5	28	52
22	4	1	3	13	21
Total........	100	89	91	111	391

* Mid-point.

† Frequency.

b. By use of formula (2), find the crude mode of the total series of ages.

2. *a.* What is the median of each of the six series below?

NUMBER OF PERSONS PER BROKEN HOME

Set I	Set II	Set III	Set IV	Set V	Set VI
3	3	3	3	3	3
5	5	5	5	2	1
4	4	4	4	4	2
1	1	2	2	1	4
6	6	6	1	5	5
8	8	8	4	8	8
2	2	2	6	11	4
11	111	11	4	5	11
				12	

b. What is the median of series IV and VI combined (added by rows)?

NOTE: These series contain too few cases for the medians to have much meaning; they are useful only for practice in finding the median.

3. *a.* Calculate the median of the two frequency distributions below:

PERCENTAGE OF CHURCHES WITHOUT A FULL-TIME MINISTER IN THE RURAL
COUNTIES OF TWO REGIONS

Percentage of churches (X^*)	Region I, counties (f_1)	Region II, counties (f_2)	Regions I and II, counties (f)
2.5	22	4	26
7.5	94	18	112
12.5	221	26	247
17.5	85	17	102
22.5	67	25	92
27.5	39	14	53
Total............	528	104	632

* Mid-point.

b. What is the median of the two distributions combined? How does it compare with the mean of the medians of the two separate distributions? What is the meaning of the mean of the medians?

4. The rural counties in 15 states were scored on various points, such as percentage of homes with telephone, per capita expenditure for

schools, and so on, and the median score for the counties in each state was determined, giving 15 medians. It was then wanted to know the median score of all counties in the 15 states together. How would you find this?

5. In the table below, what is the arithmetic mean of (*a*) the populations of the counties? (*b*) the birth rates?

County	Population (X_1)	Birth rate per 1,000 population (X_2)
1	8,003	19.5
2	21,054	24.5
3	34,301	21.1
4	15,006	9.8
5	72,573	23.1
6	15,330	16.4
7	10,233	17.4
8	16,848	12.6
9	37,581	21.2
10	34,165	16.7
11	30,503	19.1
12	16,781	21.6
13	119,217	18.3
14	52,745	14.0
15	18,182	18.6
16	46,583	16.9
17	27,037	17.3
18	42,565	22.1
19	3,815	15.5
20	59,928	16.9
21	11,471	25.2
22	38,469	19.9
23	21,953	18.1
24	13,913	13.5
25	20,039	19.2

6. Find the mean of the following table by the short method, and check it by changing the assumed mean.

Weekly wages, (X)	Women (f)
$2.50– 3.49	5
3.50– 4.49	71
4.50– 5.49	126
5.50– 6.49	132
6.50– 7.49	98
7.50– 8.49	47
8.50– 9.49	23
9.50–10.49	9
Total................	511

7. What is the mean of the table below?

Annual Net Incomes of 150 Louisiana Cotton Farms, 1936

Income . (X)*	Farms (f)†
$ 500	62
750	45
1000	23
1250	8
1500	6
1750	2
2000	2
2250	1
2500	1
Total................	150

* Mid-point.
† Frequency.

8. Calculate the mean number of years on farm reported by Iowa farmers in 1929.   Use deviations from an assumed mean.

Iowa Farm Operators Classified According to Number of Years on Farm, 1930

Years on farm (X)	Farmers (f)
Under 1 year................................	25,625
1 year..	20,140
2 to 4 years..................................	36,496
5 to 9 years..................................	33,465
10 years and over*...........................	92,142
Total..	207,868

(Abstract of the Fifteenth Census of the United States, 1930, p. 582)
* Take the mid-point of this interval at 15 years.

9. The arithmetic mean of the number of years on farm reported by **249,588** Alabama farmers in 1930 was 6.1.   What is the mean number

of years reported by Iowa and Alabama farmers combined, using the data of Exercise 8?

10. The counties of Oklahoma are to be grouped according to their infant mortality rates in 1939 as published by the Oklahoma Bureau of Vital Statistics, with the purpose of correlating these rates with the per capita expenditures for public schools. Have you any criticisms of this method?

11. A writer on the family recently made this statement: "The Census Report for 1930 showed the average size of the American family to be 3.81 persons. But averages tell us little." Can you suggest any important information that this average conceals?

12. Can you propose a refinement of the crude marriage rate analogous to the gross reproduction rate described in the text? How would it differ in meaning from the present crude rate?

13. Calculate the net reproduction rate for your state, and explain its meaning. Is the population of the state increasing or decreasing at present? If the answer to this question seems to contradict the net reproduction rate found, can you reconcile the difference?

14. What do you consider to be the most meaningful base for a divorce rate and why?

15. At what mean rate did the population of Nashville, Tenn., increase between 1880 and 1890? Between 1920 and 1930? Plot the observed populations first on rectangular coordinate paper, then on semilogarithmic paper, and study the differences.

POPULATION OF NASHVILLE, TENN., 1870–1930

Census	Population
1870	25,865
1880	43,350
1890	76,168
1900	80,865
1910	110,364
1920	118,342
1930	153,866

16. Using the data of Exercise 15, compare the geometric and arithmetic mean populations of Nashville, Tenn., between 1870 and 1930, and plot them in the graphs prepared in Exercise 15. Explain the results.

References

CHADDOCK, R. E.: *Principles and Methods of Statistics*, Chaps. VI, VII, VIII, Houghton Mifflin Company, Boston, 1925.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chap. IX, Prentice-Hall, Inc., New York, 1939.

YULE, G. U., and M. G. KENDALL: *An Introduction to the Theory of Statistics*, Chap. VII, Charles Griffin & Company, Ltd., London, 1937.

CHAPTER VIII

MEASURES OF DEVIATION AND PARTITION

1. Deviation from an Average.—It is seldom possible to give a good idea of a series of ungrouped values or of a frequency distribution by means of a single value, or average, alone. It is generally wise to exhibit the whole distribution in tabular form, and often to show it graphically as well. Mention of the range of the values, *i.e.*, the highest and lowest values in the series and the difference between them, is desirable. It is also important to accompany the average with *some measure of variation or dispersion*. The purpose of a measure of dispersion is to show the extent to which the individual items in a series vary from their average. If the average value of the items is known, and also the amount by which a certain proportion of the items deviate from that average, a rather satisfactory idea of the distribution may be conveyed. For example, note the ungrouped items 4, 1, 6, 7, 3, 9, 2, 1, 3, 4, representing the number of years between marriage and divorce in the case of 10 divorced couples. Their mean is 4 years. Six out of the 10 cases do not differ from the mean by more than 2 years. If, therefore, we describe the distribution to the reader by saying that the mean time between marriage and divorce is 4 years, and that three-fifths of the cases do not deviate from the mean by more than 2 years, he should have a better notion of the distribution than if we merely told him to imagine 10 couples whose mean time between marriage and divorce was 4 years.

2. The Average Deviation.—The simplest of the measures of dispersion is obtained by finding the amount by which each item deviates from the average value, adding these without regard to sign, and dividing the sum by the number of items, to obtain the average amount of deviation. Such a measure of deviation or dispersion is appropriately called the *average deviation*, and is often represented by the symbol $A.D.$

In the case of ungrouped data, like the above series, 4, 1, 6, 7, 3, 9, 2, 1, 3, 4, representing the number of years between mar-

riage and divorce for 10 divorced couples, the average deviation from the mean value of 4 years is found as shown in Table 33.

TABLE 33.—COMPUTATIONS FOR THE MEAN DEVIATION, UNGROUPED DATA

$$X - M_x{}^* = x$$
$$4 - 4 = 0$$
$$1 - 4 = -3$$
$$6 - 4 = +2$$
$$7 - 4 = +3$$
$$3 - 4 = -1$$
$$9 - 4 = +5$$
$$2 - 4 = -2$$
$$1 - 4 = -3$$
$$3 - 4 = -1$$
$$4 - 4 = 0$$
$$\sum_{10} |x| = 20\dagger$$

* M_x indicates the mean of the X values.

† The lines | | indicate that signs are disregarded. $\sum^{10}$ means to *add* the 10 items.

If we add the values of x with respect for the signs, the result is zero. Disregarding signs, however, the total is 20, and

$$A.D. = \tfrac{20}{10} = 2.$$

That is, the 10 values differ on the average from their mean by 2 years.

A formula for use with grouped data is

$$A.D. = \frac{\Sigma f|(X - Av)|}{N} = \frac{\Sigma f|x|}{N}. \tag{21}$$

where f is the frequency in any class interval, X is the value or mid-point corresponding to a given frequency, Av is the average used (mean, median, or mode—usually the mean), $x = X - Av$, and N is the number of items or the sum of the frequencies (f). The calculation of the $A.D.$ from the mean, M, is illustrated in Table 34. In the table, the x's are obtained, of course, by subtracting the value of the mean, 0.67, from each of the X values.

There are short methods of finding the average deviation from the mean or median, but they are rather cumbersome and will not be described here.[1]

---

[1] See, for example, H. Sorenson, *Statistics for Students of Psychology and Education*, p. 137, McGraw-Hill Book Company, Inc., New York, 1936.

TABLE 34.—NUMBER OF PREVIOUS ARRESTS RECORDED FOR 100 MURDERERS

Previous arrests (X)	Prisoners (f)	fX	x	|fx|
0	60	0	−0.67	40.20
1	20	20	+0.33	6.60
2	15	30	+1.33	19.95
3	3	9	+2.33	6.99
4	2	8	+3.33	6.66
Total............	100	67	......	80.40

$$M = \tfrac{67}{100} = 0.67.$$
$$A.D. = \frac{80.4}{100} = 0.804.$$

The average deviation is usually smaller when taken from the median than when taken from the mean or the mode.

3. The Standard Deviation.—Because the average deviation disregards negative signs, another measure of dispersion, known as the standard deviation, has been devised, which is free from this objection. It is found by subtracting each X value, or in grouped data each mid-point value, from the mean of the X values, squaring these differences to make all signs positive, multiplying them by their respective frequencies, summing them, dividing by the sums of the frequencies, and extracting the square root. The formula for the standard deviation is, therefore,

$$\sigma = \sqrt{\frac{\Sigma(X - M_x)^2}{N}}, \qquad (22)*$$

or

$$\sigma = \sqrt{\frac{\Sigma f(X - M_x)^2}{N}}. \qquad (23)$$

Letting $x = X - M_x$,

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}, \qquad (24)$$

or

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}}, \qquad (25)$$

* The Greek letter, small sigma, σ, is conventionally used to represent the standard deviation.

where σ is the standard deviation of the X values, X is the value of an item or the value of the mid-point of a group of items, f is the frequency of the items in a group or class interval (for ungrouped data, $f = 1$), and N is the number of items, *i.e.*, $N = \Sigma f$.

To save labor in computing the standard deviation for a large frequency table, a short method is commonly used:

$$\sigma = i \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2}, \qquad (26)[1]$$

where d is the deviation of the mid-points from a guessed mean in class interval units, i = width of class interval. This formula may also be modified for use with ungrouped data by taking the assumed[2] mean at zero, so that $d = X, f = 1$, and $i = 1$:

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}, \qquad (27)$$

or

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - M_x{}^2}. \qquad (28)$$

---

[1] Derivation of formula (26):

By definition,

$$\sigma = \sqrt{\frac{\Sigma f(X - M_x)^2}{N}}. \qquad (23)$$

From Chap. VII, formulas (10) and (13),

$$X = A + id \qquad (a)$$

$$M = A + \frac{i\Sigma f d}{N}. \qquad (b)$$

Substituting from (a) and (b) in (23)

$$\sigma = \sqrt{\frac{1}{N} \Sigma f \left(A + id - A - \frac{i\Sigma f d}{N}\right)^2}, \qquad (c)$$

$$\sigma = \sqrt{\frac{1}{N} \Sigma f \left(id - \frac{i\Sigma f d}{N}\right)^2},$$

$$\sigma = \sqrt{\frac{1}{N} \Sigma f \left[i^2 d^2 - 2i^2 d \frac{\Sigma f d}{N} + i^2 \left(\frac{\Sigma f d}{N}\right)^2\right]},$$

$$\sigma = i \sqrt{\frac{\Sigma f d^2}{N} - 2\left(\frac{\Sigma f d}{N}\right)^2 + \frac{\Sigma f}{N}\left(\frac{\Sigma f d}{N}\right)^2},$$

$$\sigma = i \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2}. \qquad (26)$$

[2] See Chap. VII.

In Chap. VII, we had the ungrouped data, 3, 7, 2, 12, 1, 16, 4, representing the numbers of children in seven Italian immigrant families. The mean number of children per family was found to be 6.43. What is the standard deviation? If we use the long method of formula (22) or (24) above, we require the computations shown in Table 35.

TABLE 35.—COMPUTATIONS FOR THE STANDARD DEVIATION, UNGROUPED DATA
(Long method)

X	$X - M_x$	$(X - M_x)^2$
3	$3-6.43 = -3.43$	$(-3.43)^2 = 11.76$
7	$7-6.43 = +0.57$	$(0.57)^2 = 0.32$
2	$2-6.43 = -4.43$	$(-4.43)^2 = 19.62$
12	$12-6.43 = +5.57$	$(5.57)^2 = 31.02$
1	$1-6.43 = -5.43$	$(-5.43)^2 = 29.49$
16	$16-6.43 = +9.57$	$(9.57)^2 = 91.58$
4	$4-6.43 = -2.43$	$(-2.43)^2 = 5.90$
Total........	....	189.69

Substituting in formula (22),

$$\sigma = \sqrt{\frac{189.69}{7}} = 5.21.$$

For the short method of formula (27) or (28), we need only the two totals, as shown in Table 36.

TABLE 36.—COMPUTATIONS FOR THE STANDARD DEVIATION, UNGROUPED DATA
(Short method)

X	X^2
3	9
7	49
2	4
12	144
1	1
16	256
4	16
45	479

Substituting in formula (27),

$$\sigma = \sqrt{\tfrac{479}{7} - (\tfrac{45}{7})^2},$$
$$\sigma = 5.21,$$

as before. The saving of labor in comparison with the first method is evident.

Let us next find the standard deviation of Table 34 above, and compare it with the average deviation previously obtained for the same table. We shall again first employ the long method, to clarify the meaning of the arithmetic, and to enable the student to compare the amount of work required relative to the short method to follow. The formula that describes the long method for grouped data is formula (23) or (25), which calls for the computations shown in Table 37. The mean of the table is 0.67.

TABLE 37.—COMPUTATION OF STANDARD DEVIATION FOR TABLE 34
(Long method)

X	f	$X - M_x = x$	$f(X - M_x) = fx$	$f(X - M_x)^2 = fx^2$
0	60	-0.67	-40.20	26.93
1	20	$+0.33$	$+ 6.60$	2.18
2	15	$+1.33$	$+19.95$	26.53
3	3	$+2.33$	$+ 6.99$	16.29
4	2	$+3.33$	$+ 6.66$	22.18
Total.........	100		0.00	94.11

Substituting in formula (23) or (25),

$$\sigma = \sqrt{\frac{94.11}{100}} = 0.97.$$

Turning now to the short method of formula (26), the steps are worked out in Table 38. Notice the so-called *Charlier check*

TABLE 38.—COMPUTATION OF STANDARD DEVIATION FOR TABLE 34
(Short method)

X	f	d	fd	fd^2	$f(d + 1)^2$
0	60	-1	-60	60	0
1	20	0	0	0	20
2	15	$+1$	$+15$	15	60
3	3	$+2$	$+ 6$	12	27
4	2	$+3$	$+ 6$	18	32
Total.........	100		-33	105	139

included in Table 38: $\Sigma f + 2\Sigma fd + \Sigma fd^2 = \Sigma f(d+1)^2$, or $100 + 2(-33) + 105 = 139$, which is the total of the last column of the table. This checks all of the work of the table. Substitution in formula (26) now gives

$$\sigma = 1\sqrt{\frac{105}{100} - \left(\frac{-33}{100}\right)^2},$$
$$\sigma = 0.97,$$

which is the value reached by the long method.[1]

The average deviation of Table 34 was found in Sec. 2 above to be 0.804, while we see that the standard deviation is 0.97. The standard deviation is always larger than the average deviation, because squaring the differences gives greater weight to the extreme values.

Because of the inaccuracies due to grouping data in class intervals, the standard deviation squared, called the *variance*, of a distribution that is fairly symmetrical[2] in form is often corrected by subtracting from it the value $i^2/12$ in the case of a continuous variable, or $\left(\frac{i^2}{12} - \frac{1}{12}\right)$ in the case of a discrete variable. In the above problem the variable is discrete, so that we have $(0.97)^2 - (\frac{1}{12} - \frac{1}{12}) = (0.97)^2$, and σ remains unchanged. There is no error of grouping when the variable is discrete and $i = 1$. This correction is known as *Sheppard's correction*. In its usual form it cannot be applied to very skewed or asymmetrical distributions.

If we have calculated the standard deviation of each of two series, and then wish to know the standard deviation of the two series combined, the latter may be found from the formula

$$\sigma = \sqrt{\frac{N_1(\sigma_1^2 + M_1^2) + N_2(\sigma_2^2 + M_2^2)}{N} - M^2}, \qquad (29)$$

where the subscripts differentiate the two series, and no subscript indicates the combined series. Where there are more than

---

[1] In Table 38, it happens that the X values are already in unit step deviation form—0, 1, 2, etc.—so that very little labor is saved by using the d column. We might, therefore, have used X in place of d in formula (26). The student is asked to do this as a check on the calculations in Table 38.

[2] The distribution should be normal in form. See Chap. IX.

two series, a term $N_i(\sigma_i{}^2 + M_i{}^2)$ is inserted in the formula for each additional series.

For example, for Table 34 we have $N_1 = 100$, $\sigma_1{}^2 = 0.94$, and $M_1{}^2 = 0.45$. In a second sample of the same kind, given $N_2 = 80$, $\sigma_2{}^2 = .6302$, and $M_2{}^2 = 4.56$. From formula (29), for the two samples combined, we find

$$\sigma = \sqrt{\frac{100(.94 + .45) + 80(.6302 + 4.56)}{180}} - 1.74,$$

$$\sigma = 1.16.$$

Just as the average deviation is usually a minimum when taken from the median, so the standard deviation is a minimum when taken from the mean. In fact, the standard deviation is practically never taken from any average except the mean, and formulas (27) and (28), above, are valid only for the mean.

4. Effect of Coding[1] on Averages and Measures of Dispersion. If the frequencies in a frequency table are divided through by a constant, k, the averages and measures of dispersion or partition calculated from the table will not be changed. Since it is possible to simplify the computation in this way, it is desirable to use this device whenever the opportunity offers.

The student is asked to test this for himself, using Table 39, in calculating the mean and the standard deviation.

TABLE 39.—MEAN ANNUAL INCOME OF 500 CLERICAL WORKERS

Mean Income (X)	Families (f)
$ 500	25
1,000	150
1,500	200
2,000	75
2,500	50
Total..................	500

It is also often convenient to reduce the absolute frequencies to percentage frequencies before using them in computation.

5. The Coefficient of Variation.—The average or standard deviations of two frequency distributions are not directly comparable, because they depend upon the size of the mean or median in each case, and upon the particular unit used. For example, the weights of a herd of elephants may vary on the

---

[1] Dividing the frequencies of a distribution by a constant.

average by 280 lb., while the weights of a litter of mice may differ by 0.1 oz. Yet the mice may show a greater variation than the elephants relative to their mean weights. Average and standard deviations may, therefore, be made comparable by expressing them as percentages of their means or medians. This percentage is called the *coefficient of variation* in terms of the average or standard deviation, and is written

$$V = \frac{100 A.D.}{M},\qquad(30)$$

or

$$V = \frac{100 A.D.}{Md},\qquad(31)*$$

and

$$V = \frac{100\sigma}{M}.\qquad(32)$$

It is possible to use the coefficient of variation, V, as a measure of the representativeness of an average. It may be said, arbitrarily, that when V is above 50 per cent, it is usually advisable to abandon the use of an average as a single value intended to give an idea of the central tendency of a series. The V calculated for the mean of Table 34 above by formula (30) is

$$V = \frac{100(0.804)}{0.67} = 120 \text{ per cent.}$$

In this case, V is 70 points above 50 per cent; hence the mean is obviously a poor device for representing the actual values in this very skewed or J-shaped distribution. If we apply formula (30) to the mean of Table 40, below, which is merely a rearrangement

TABLE 40.—PREVIOUS ARRESTS RECORDED FOR 100 MURDERERS
(Frequencies of Table 36 rearranged)

X	f
0	2
1	20
2	60
3	15
4	3
Total..................	100

$$M = 1.97,$$
$$A.D. = 0.467.$$

* Only one of these formulas should be used in the same comparison.

of the frequencies of Table 34 in more symmetrical form for purposes of illustration, we find that $V = 100(0.467)/1.97 = 24$ per cent, indicating that the mean represents the values in this table very well. This result would be expected from an inspection of the distribution, which appears to be fairly symmetrical in form, with the largest frequency in the center.

In using formulas (30), (31), and (32), it will be seen that if two distributions have equal average or standard deviations, but unequal means or medians, the one with the larger average will have the smaller coefficient of variation, V. This is as it should be, provided that the means or medians used in finding the V's contain no element that spuriously raises or lowers the values from which the averages are calculated.

Suppose that the question is asked, Does Table 34 or Table 40 show a greater amount of variability from the mean? In the case of Table 34 it has been seen that $V = 120$ per cent, and for Table 40 it was found that $V = 24$. The V's, therefore, show that Table 34 is $\frac{120}{24} = 5$ times as variable as Table 40, whereas the average deviations would indicate that the former distribution was less than twice as variable as the latter.

6. Partition Values.—To show the scale values below which any desired proportion of the frequencies in a distribution fall, a set of partition values known as *quartiles, deciles,* etc., or more inclusively as *percentiles,* has been devised. These measures all employ the principle of the median, and apply primarily to grouped data. Thus, while the median is that scale value below which half of the values fall, the first quartile, Q_1, is the scale value below which lie one-fourth of the values; the third quartile, Q_3, is the scale value below which lie three-fourths of the values; the ninth decile, d_9, is the scale value below which lie 90 per cent of the values; the 65th percentile is the scale value below which lie 65 per cent of the values; and so on. It is, therefore, seen that each of these measures is merely a particular percentile value, the median corresponding to the 50th percentile, the first quartile to the 25th percentile, the third quartile to the 75th percentile. The general method of finding any value is the same.

Because of logical difficulties, it is seldom that any partition value except the median is found for ungrouped data.[1]

[1] If the attempt must be made, however, it is generally best to accept rough approximations, rather than insist on exact but imaginary interpola-

For grouped data, it will be recalled that the median is located by dividing the total frequency, N, by 2, counting up the column of accumulated frequencies of the table until the lower limit of the class interval is reached which contains the median value, and then interpolating within this interval to determine the median value. When finding any percentile value other than the median, we need only change the coefficient of the total

---

tions. For example, if we are required to furnish the third quartile, Q_3, for the array of 12 ages—3, 5, 6, 9, 11, 16, 20, 21, 24, 25, 26, and 30 years— we may find the position $12 \times 0.75 = 9$, and say that $100 \left(\frac{9}{12}\right) = 75$ per cent of the ages are *less than* the age of 25 years that occupies $9 + 1 = 10$th place in the array. This statement is correct in the present case; but it is not correct to say, further, that $100 - 75 = 25$ per cent of the ages are *greater than* 25 years. If the age 24 years in the array were replaced by a second age 25 years, then the age 25 years would no longer be greater than 75 per cent of the ages, but it would still probably be the most appropriate age to offer as an approximate value for Q_3.

When the position, Np, found by multiplying the total number of items, N, by the given percentage value, p, is not a whole number, the matter is more complicated. Thus, if we drop the age 30 years from the top of the above array, we have $Np = 11 \times .75 = 8.25$. There is no 8.25th position in this array, so we have to choose between positions number 8 and 9, or else interpolate between them. If we take position 8 as the nearest integer, and add one to it, as we did above, we get position 9. The age corresponding to this position is 24 years, and we see that eight ages, or $100\left(\frac{8}{11}\right) = 72.7$ per cent of the ages, are less than this age. Since 72.7 per cent is rather close to 75 per cent, the age 24 years seems to be the simplest approximate value to assign to Q_3.

Only when no actual position in an array gives a reasonably close approximation to the meaning of a required percentile is it usually worth while to interpolate between two positions. If our array above consisted of only the first 10 ages, to find Q_3 we would have $pN = 0.75(10) = 7.5$. The age in the eighth position is greater than $100\left(\frac{7}{10}\right) = 70$ per cent of all the ages, whereas that in the ninth position is greater than $100\left(\frac{8}{10}\right) = 80$ per cent of the ages. Here we might prefer to take the interpolated position, $\frac{7 + 8}{2} = 7.5$, so that, *assuming continuous or grouped data*, the theoretical age corresponding to it would be greater than $100(7.5/10) = 75$ per cent of the ages in the array. This theoretical age, or value of Q_3, must be halfway between age 20 in seventh position and age 21 in eighth position, or $\frac{20 + 21}{2} = 20.5$ years.

Notice that, in ungrouped data, the empirical formula, $Np + 1$, used for locating the approximate integral position of such a partition value as Q_3, is replaced by the formula $p(N + 1)$ for determining the median position.

frequency, N. For example, in the case of Q_1, we use $N/4$, for
Q_3, $3N/4$, for d_9, $0.9N$, for the 65th percentile, $0.65N$, and so on.
The general formula, using P to represent any percentile, median,
decile, or quartile value on the X scale, is

$$P = L + \left(\frac{pN - F}{f}\right) i, \qquad (33)$$

where p is the percentile rank or point of division on the *frequency*
scale expressed in percentage form (*e.g.*, $p = 0.75$), L is the lower
limit of the interval containing the pth value, N is the total
frequency of the table, F is the sum of the frequencies falling
below (*i.e.*, in class intervals with limits smaller than) L, f is
the number of frequencies in the interval containing the pth
value, and i is the size of interval containing the pth value.

Let us find the values of Q_1, Q_3, d_7, and p_{33} (33rd percentile)
in Table 41.

TABLE 41.—DISTRIBUTION OF THE ESTIMATED INCOME AMONG UNMARRIED
WOMEN OF THE UNITED STATES IN 1910*

Income (X)	Women (Y)	(Y) Accumulated	Percentage accumulated
$ 100– 199	10	10	0.55
200– 299	70	80	4.42
300– 399	560	640	35.36
400– 499	530	1,170	64.64
500– 599	280	1,450	80.11
600– 699	150	1,600	88.40
700– 799	110	1,710	94.48
800– 899	37	1,747	96.52
900– 999	22	1,769	97.73
1,000–1,099	16	1,785	98.62
1,100–1,199	12	1,797	99.28
1,200–1,299	8	1,805	99.72
1,300–1,399	5	1,810	100.00
Total.............	1,810		

* From W. I. KING, *Wealth and Income of the People of the United States*, p. 224, 1915.

To find Q_1, we have

$$pN = .25(1,810) = 452.5.$$

Counting up (*i.e.*, in the direction of increasing values on the X
scale) the accumulated frequency column of the table, we see that

452.5 lies in the class interval 300–399.   Therefore,

$$L = 300.$$
$$F = 80.$$
$$f = 560.$$
$$i = 100.$$

Substituting in equation (33),

$$Q_1 = 300 + \frac{452.5 - 80}{560} \cdot 100,$$

$$Q_1 = 366.5.$$

That is, one-fourth of the women earned less than $366.50 a year. Similarly,

$$Q_3 = 500 + \frac{1,357.5 - 1,170}{280} \cdot 100,$$

or

$$Q_3 = 567,$$
$$d_7 = 500 + \frac{1,267 - 1,170}{280} \cdot 100,$$

or

$$d_7 = 534.6,$$
$$P_{33} = 300 + \frac{597.3 - 80}{560} \cdot 100,$$

or

$$P_{33} = 392.4.$$

From these results we notice that three-fourths of the working women made less than $567 annually, 70 per cent of them made below $534.60, and one-third made under $392.40.   Of course, there is no point in calculating all of these values except for illustrative purposes.   We are usually interested in such fractions as one-third, one-half, or three-fourths.

An investigator often requires, not the value below which a certain percentage of the frequencies fall, but the reverse of this, namely, the percentage of the cases that falls below a certain value, that is, the *percentile rank* of the value.   Referring back to the ungrouped array of 11 ages used above, *viz.*, 3, 5, 6, 9, 11, 16, 20, 21, 24, 25, and 26 years, we may require the percentile rank of the person aged 21.   Since, by definition, this is equivalent to asking what percentage of the persons in the array are less than

21 years of age, we note that there are 7 persons out of 11 who are younger than 21 years, and compute $\frac{7}{11} = 0.636$, or 63.6 per cent. We then say that the percentile rank of the person aged 21 is approximately 64.

Turning to grouped data, suppose we ask what proportion of the unmarried women of Table 41 earned less than some minimum living wage, say \$550 a year. Our problem now is, knowing a value on the X scale, to find the percentage of values on the Y scale that fall below it. In the present case, it is evident that 1,170 women earned less than \$500, and that 280 earned between \$500 and \$599. We have $\dfrac{550 - 500}{600 - 500}(280) = 140$, as the number of women earning between \$500 and \$550. Therefore $1,170 + 140 = 1,310$ is the number of women who made less than \$550. Expressed as a percentage of the total number of women workers, we find that $100\left(\dfrac{1,310}{1,810}\right) = 72$ per cent of the women failed to earn as much as the minimum amount. A formula for this calculation is

$$p = \left[F + \frac{f(P - L)}{i} \right] \frac{100}{N}, \tag{34}$$

where p is the percentile rank sought, P is the given X scale value, F is the accumulated frequencies in the class intervals with limits smaller than those of the interval including P, f is the frequency of the interval including P, L is the lower limit of this same interval, i is its width, and N is the total frequency of the table. Thus, substituting the values of the preceding problem in formula (34), we get

$$p = \left[1,170 + \frac{280(550 - 500)}{100} \right] \frac{100}{1,810},$$

or $p = 72$ per cent, as before.

An X scale value corresponding to a given accumulated frequency, or a percentage frequency corresponding to a given X scale value, may also readily be found by means of a cumulative curve, which was described in Fig. 11, Chap. VI. The student is asked to use this device to check the arithmetical results just obtained from Table 41 above, preferably plotting the curve from the last column of that table.

A measure known as the *quartile deviation* is sometimes used. The formula is

$$Q = \frac{Q_3 - Q_1}{2}. \tag{35}$$

Thus, for Table 41,

$$Q = \frac{567 - 366.5}{2} = 100.25.$$

The quartile deviation is employed only when the median is the preferred average.

All these measures of dispersion—quartiles, deciles, percentiles,
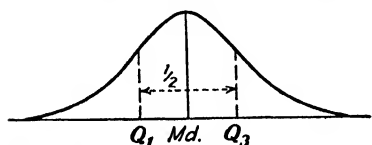
Fig. 37.—The distance Q_3-Q_1 includes half of the cases.

quartile deviation—are so-called *position* values, and have the same advantages and disadvantages as the median, previously discussed. In particular, they are insensitive to extreme values, and cannot be treated algebraically. They are especially useful in analyzing a skewed frequency distribution, since they maintain a definite relationship to the distribution, regardless of its shape.

7. Comparable Measures or Scores.—When two frequency distributions are of about the same shape, *e.g.*, both about symmetrical, both slightly skewed in the same direction, both J-shaped, etc., distances on their scales are usually compared in units of their respective standard deviations. Thus, if we have the distributions of many scores on two independent tests of a given trait, for each test the deviations of the scores from the true[1] mean are divided by the true standard deviation, to get the desired *standard scores*. Given, for Test I, true mean = 70, true σ = 10; and for Test II, true mean = 62, true σ = 12. If subject A scored 80 on Test I and 60 on Test II, his standard score on Test I is $\frac{80 - 70}{10} = 1$, and on Test II is $\frac{60 - 62}{12} = -0.17$; and his combined score on the two tests is $1 + (-.17) = 0.83$. If subject B scored 75 on Test I and 65 on Test II, his corresponding standard scores are $\frac{75 - 70}{10} = 0.50$

[1] By *true* is meant a statistic derived from many applications of a test, rather than from a single application, to the same *universe* or type of subjects.

on Test I, and $\frac{65 - 62}{12} = 0.25$ on Test II; and his combined score is 0.75.

Where two distributions differ markedly in form, *e.g.*, one being about symmetrical and the other J-shaped, or one very peaked and the other flat, the standard deviations do not provide consistent units for reducing their scale distances to more comparable terms, because the proportion of frequencies included between the mean and one standard deviation on each side of it changes with the form of the distribution. Theoretically, perhaps the best procedure under these circumstances is to *normalize* both distributions, but the method is too complex to introduce here.[1] A cruder but much simpler method uses the Q's instead of the σ's as common denominators. Although Q also has disadvantages, it is one-half of the range Q_3-Q_1, within which always falls the middle half of the frequencies; and in that sense its interpretation is independent of the shape of the distribution (see Fig. 37).

Suppose now that the distribution of many scores in Tests I and II above are quite different, being J-shaped to the left in Test I and skewed to the right in Test II. For Test I the true median score is 74, and the true Q value is 6; for Test II, the median score is 59 and Q is 8. We divide the deviations of the two subjects' scores from the medians by the respective Q values, and get $\frac{80 - 74}{6} + \frac{60 - 59}{8} = 1.125$ as the combined score of subject A, and $\frac{75 - 74}{6} + \frac{65 - 59}{8} = 0.917$ as the combined score of subject B. These may be called the Q scores.

Instead of the standard scores or Q scores described above, the method of equivalent percentile scores may be used in the effort to make two independent scales comparable. For each scale, every percentile or, say, every fifth percentile is found, and these values are arranged in two parallel series, where corresponding pairs of values are regarded as equivalent. Thus, in Table 42 below, the values $X_1 = 13$ and $X_2 = 0.1$, are equivalent on the two scales. The percentile values are found arithmetically from the two given frequency distributions of scale values by formula (33)

[1] PAUL HORST, Obtaining Comparable Scores from Distributions of Dissimilar Shape, *Journal of the American Statistical Association*, Vol. 26, pp. 455–460, 1931.

above, or graphically from the ogive curve as illustrated in Fig. 11, Chap. VI.   Suppose that we wish to compare the score of subject 114 on Test X_1, 85, with the score of subject 17 on Test X_2, 2.6.   From Table 42 we see that a score of 2.6 on scale X_2 is equivalent to a score of 93 on scale X_1.   Hence the two comparable scores are 85 and 93.   If either or both of the scores of subjects 114 and 17 did not appear in Table 42, we would find the percentile rank of say the second of them by formula (34) above, and then, using this in formula (33), find the corresponding value on the X_1 scale.   This equivalent X_1 value would then be compared with the X_1 score of the other subject.

TABLE 42.—TWO SERIES OF EQUIVALENT PERCENTILE SCALE VALUES: ATTITUDE TOWARD WAR

n	Scale, X_1 (P_n)*	Scale, X_2 (P_n)
5	13	.1
10	23	.4
15	32	.5
20	41	.6
25	49	.65
30	56	.8
35	63	1.0
40	69	1.2
45	75	1.4
50	80	1.6
55	85	1.9
60	89	2.2
65	93	2.6
70	95	2.9
75	97	3.2
80	97.5	3.6
85	98	3.9
90	98.5	4.2
95	99	4.6
100	100	5.0

* nth percentile scale value.

The above three methods are not applicable to ungrouped or scanty data.

When the data are inadequate, or when for other reasons we have more confidence in the ability of two scales to arrange items

in rank order than to measure distances between them, simple percentile ranks may be used for purposes of comparison. Given the scores on a test, the percentile rank is found for each score. For example, if 62 per cent of the scores made on a test are less than the score 80, the percentile rank of the latter is 62. For ungrouped data, the percentile ranks are found by the informal method outlined on page 132; for grouped data, the percentile ranks are obtained arithmetically from formula (34), above, or graphically from an ogive. The weakness of percentile ranks is, of course, that they do not reflect the distances between the scores on any scale. Thus, the score 70 may have a percentile rank of 50, the score 77 a percentile rank of 60, and the score 85 a percentile rank of 90, so that the successive scores stand in the ratio of $1:1.1$, whereas the corresponding successive percentile ranks bear the ratios $1:1.2$ and $1:1.5$, respectively. For this reason, the difference between percentile ranks should not be interpreted as proportional to the distance between the corresponding scale values.

As a matter of fact, there is usually no feasible method of treating scores obtained from the use of very different kinds of scales that makes them strictly comparable.

Exercises

1. Compare the average deviation and the standard deviation of the series below. Find the standard deviation by formulas (24) and (28) as a check.

NUMBER OF DEPENDENTS IN 25 FAMILIES ON RELIEF

Family no.	Dependents	Family no.	Dependents
1	3	14	5
2	5	15	3
3	4	16	3
4	1	17	2
5	6	18	4
6	8	19	1
7	2	20	3
8	3	21	4
9	3	22	3
10	2	23	6
11	4	24	2
12	1	25	3
13	2		

2. Compare the average deviation and the standard deviation of the following frequency distribution, using for the standard deviation formula (26) with the Charlier check:

SEMESTER HOURS OF MATHEMATICS TAKEN BY 67 STUDENTS IN A CLASS OF
ELEMENTARY SOCIAL STATISTICS

Semester Hours	Students
43.5–46.4	1
40.5–43.4	0
37.5–40.4	0
34.5–37.4	0
31.5–34.4	0
28.5–31.4	2
25.5–28.4	2
22.5–25.4	5
19.5–22.4	4
16.5–19.4	8
13.5–16.4	13
10.5–13.4	26
7.5–10.4	4
4.5– 7.4	1
Total..................	66

3. Use the coefficient of variation, V, to measure the representativeness of the mean of the distribution in Exercise 2, above.

4. Below are two random samples of family incomes in a certain city, one taken in 1928, the other in 1932. Did the depression reduce or increase the spread in income between families?

Income	Number of families	
	1928	1932
Under $500.........................	5	76
500–999.............................	15	123
1,000–1,499.........................	115	155
1,500–1,999.........................	190	91
2,000–2,499.........................	82	70
2,500–2,999.........................	63	52
3,000–3,499.........................	27	17
3,500–3,999.........................	19	12
4,000–4,499.........................	10	7
4,500–4,999.........................	6	3
5,000–5,499.........................	3	1
Total.............................	535	607

5. Using the standard deviations found for the 1928 and 1932 series in Exercise 4, compute the standard deviation for the two series combined.

6. The table below shows the number of children who required the specified numbers of hours of social contact before they were "accepted" in a certain play group. (*a*) What percentage of the children took less than 4 hours? (*b*) What percentage of the children took more than 10 hours? (*c*) How many hours did three-fourths of the children require less than? (*d*) How many hours did three-fourths of the children require more than?

Hours	Children
18–19	1
16–17	3
14–15	2
12–13	6
10–11	10
8– 9	9
6– 7	8
4– 5	6
2– 3	3
0– 1	2
Total...................	50

7. Given two independent scales, X and Y, for the measurement of "cooperation" between members of a random sample of urban families. Family A has a score of $+1.2$ on scale X, family B has a score of 86 on scale Y. Reduce these scores to as nearly comparable terms as you can.

Scale X	Families	Scale Y	Families
-2.5–-2.9	4	0– 9	21
-2.0–-2.4	12	10–19	68
-1.5–-1.9	22	20–29	109
-1.0–-1.4	45	30–39	140
-0.5–-0.9	71	40–49	131
0.0–-0.4	89	50–59	91
0.0–$+0.4$	116	60–69	74
$+0.5$–$+0.9$	132	70–79	56
$+1.0$–$+1.4$	151	80–89	28
$+1.5$–$+1.9$	93	90–99	13
$+2.0$–$+2.4$	60	Total.............	731
$+2.5$–$+2.9$	17		
Total..............	812		

8. Suppose that the frequencies for the Y scale in Exercise 7, are reversed end for end of the scale, while those for the X scale remain as they are.   Convert these scores to a more comparable basis.

References

CHADDOCK, R. E.: *Principles and Methods of Statistics*, Chap. IX, Houghton Mifflin Company, Boston, 1925.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chap. X, Prentice-Hall, Inc., New York, 1939.

DAVIES, G. R., and DALE YODER: *Business Statistics*, Chap. II, John Wiley & Sons, Inc., New York, 1937.

GARRETT, H. E.: *Statistics in Psychology and Education*, Chap. 1, Longmans, Green & Company, New York, 1926.

KELLEY, T. L.: *Statistical Method*, Chap. VI, The Macmillan Company, New York, 1923.

LINDQUIST, E. F.: *A First Course in Statistics*, Chap. IX, Houghton Mifflin Company, Boston, 1938.

MILLS, F. C.: *Statistical Methods*, rev. ed., Chap. V, Henry Holt and Company, Inc., New York, 1938.

SORENSON, H.: *Statistics for Students of Psychology and Education*, Chaps. VII and VIII, McGraw-Hill Book Company, Inc., New York, 1936.

WHITE, R. C.: *Social Statistics*, Chap. IX, Harper & Brothers, New York, 1933.

CHAPTER IX

COMBINATION, PROBABILITY, AND THE NORMAL DISTRIBUTION

1. Permutations and Combinations.[1]—It is often desirable in sociological investigations to know the total number of ways in which a certain event can occur. For example, in a study of intercity migration among five cities, how many paths can the migration take? Or, among 10 girls in a boarding school, three two-girl friendships are found. How many such friendships are possible in this group? The same kind of problem arises in connection with the binomial formula, discussed in Sec. 3 below.

To answer the question about the paths of migration, we notice that since a migrant may go from any of the five cities to any of the four remaining cities, the number of paths must be $5 \times 4 = 20$. Not only do we count each pair of cities, but also the two orders or arrangements in which the members of a pair may be taken, as "from a to b," and "from b to a." A pair of cities in a given order, *e.g.*, "from a to b," is called a *permutation*, and the general formula provided by algebra for finding the number of permutations of n things taken r at a time is

$$_nP_r = \frac{n!}{(n-r)!}. \tag{36}[2]$$

For the problem above, we substitute in the formula, and get

$$_5P_2 = \frac{5!}{(5-2)!} = \frac{(5 \times 4 \times 3!)}{3!}$$
$$= 5 \times 4 = 20,$$

as before.

Formula (36) is based on Theorem 1.

---

[1] For a fuller treatment of this subject, see any text in college algebra, *e.g.*, H. B. Fine, *College Algebra*, Chap. XXV, Ginn and Company, Boston, 1904.

[2] $n!$ is called "n factorial," and means the product of all consecutive numbers from 1 through n. For example, $4! = 4 \times 3 \times 2 \times 1 = 24$.

THEOREM 1.   *If an event A can occur in m ways, and thereafter an event B can occur in n ways, A and B can occur together in the order named in mn ways.*

A first approach to the problem of the boarding school friendships mentioned above can also be made by means of formula (36).   The number of arrangements, or permutations, of 10 girls taken two at a time is

$$_{10}P_2 = \frac{10!}{8!} = 10 \times 9 = 90$$

Here, however, there is no interest in the order of the girls in a two-girl friendship.   When this is the case, *i.e.*, when a group of things is taken without regard for the arrangement of the members, the group is called a *combination*.   Evidently, each pair of girls can be arranged in two orders or permutations, so that the 90 permutations found above reduce to $\frac{90}{2} = 45$ combinations.   The formula for combinations is, therefore,

$$_nC_r = \frac{_nP_r}{r!} = \frac{n!}{r!(n-r)!}. \tag{37}$$

Using it, we get again

$$_{10}C_2 = \frac{10!}{2!\,8!} = \frac{10 \times 9 \times 8!}{2!\,8!}$$
$$= \frac{10 \times 9}{2 \times 1} = \frac{90}{2} = 45.$$

Although formulas (36) and (37) apply to a large number of problems, some problems occur that are best approached independently.   As an easy example, suppose we ask, What is the total number of possible relationships that can exist between two persons, X and Y, in terms of attraction, indifference, and repulsion?   To each of the three attitudes of X, Y may respond with three attitudes, so that, by Theorem 1 above, we have $3 \times 3 = 9$ relationships.   These relationships are (1) mutual attraction between X and Y; (2) X is attracted by Y, but Y is indifferent to X; (3) X is attracted by Y, but Y is repulsed by X; (4) Mutual indifference between X and Y; (5) X is indifferent to Y, but Y is attracted by X; (6) X is indifferent to Y, but Y is repulsed by X; (7) mutual repulsion between X and Y; (8) X is repulsed by Y, but Y is indifferent to X; and (9) X is repulsed by Y, but Y is attracted by X.

2. Probability.—Chance, often called "luck," and the tricks it plays are known to everyone. In a hand at cards, one may draw no ace; one, two, three, or even all four aces. Whether a person is male or female, white or black, European or American, is, as far as he is concerned, purely an accident. The occupation one follows, the person one marries, the state of one's health, and so on, are also subject to a great amount of chance. Discovery and invention, even the trend in the development of a nation's culture in the sociological sense, depend in part on thousands of small forces of which we have no knowledge. If the birth rates in a city differ in 1939 and 1940, is it because fundamental conditions affecting fertility have changed, or is the variation due merely to accidental factors that will cancel out over several years? In one random sample of old people there may be more male than female survivors and in another sample exactly the reverse, regardless of the true proportion in the population. It is, therefore, not surprising that any careful attempt to investigate social life or culture is obliged to reckon with this element of chance. Chance distorts the findings of research, and must be allowed for.

One of the greatest practical contributions of mathematics has been its discovery, beneath apparent confusion, of a remarkable regularity in the occurrence of chance events. By mathematical means, we can estimate the amount of variation due to chance and predict the number of occurrences of any event whose probability is known, *e.g.*, the annual deaths in a class of insurance risks. On these mathematical laws of probability are founded great business enterprises like insurance, as well as the basic techniques of a vast amount of scientific and industrial research.

The exact *mathematical definition of probability* is this: If an event can succeed in m ways and fail in m' ways, all equally likely and mutually exclusive, and the event must either succeed or fail, the probability of its succeeding is

$$p = \frac{m}{m + m'},$$ (38)

and that of its failing is

$$q = \frac{m'}{m + m'}.$$ (39)

That is,

$$p + q = \frac{m + m'}{m + m'} = \frac{1}{1} = 1. \tag{40}$$

In other words, since an event must either succeed or fail, the probability of certainty is one in one, or unity.

The proportion of ways in which an event can succeed may be determined for practical purposes by one of two methods, or by both. In the case of a penny, we decide that the probability of throwing a head is $\frac{1}{2}$, by reasoning that the penny has only two sides and is equally balanced so that one of them is as likely to turn up as the other. This is an illustration of the *theoretical* or *a priori* method. By the so-called *empirical* method, the chance of death within a year of a white male, aged 30, engaged in a clerical occupation, married, and an "A" medical risk, is found by simply counting the proportion of annual deaths occurring among a very large number of such individuals (say, 354 deaths among 85,707 persons, giving a probability of 0.00413). The empirical method is sound if the probability tends to approach a limit, as the estimate is based on an ever-increasing number of cases under essentially the same conditions. In both methods, of course, it is supposed that the conditions under which the probability was obtained will hold approximately for all situations to which the probability is applied. For example, if each added count of deaths in a risk group like that described above causes the average probability of death to approach nearer to some figure 0.00400, then 0.00400 may be regarded as an approximation of the true (expected) proportion that exists in the given class as a whole (an infinite universe). But it would obviously be wrong to apply this death rate to a class in which the age was 40 instead of 30 years!

Two basic theorems of probability are

THEOREM 2. *Of two mutually exclusive[1] events, A and B, if the event A has a probability of occurring, p, and the event B has a probability of occurring, p', the probability that either A or B will occur in one possible way is p + p'.*

[1] Two events are mutually exclusive when in a single trial only one of them can happen. In a hand at cards, drawing an ace and drawing a jack are mutually exclusive events, but drawing an ace and drawing a diamond are not, because both may appear on the same card. If the two events are not mutually exclusive, the probability is $p + p' - pp'$.

THEOREM 3.   *If an event A has a probability of occurring, p, and an event B has a probability of occurring with or after A in one possible way, p', the probability that both A and B will so occur is pp'.*

A first application of Theorem 3 may be made to a typical problem.   A community is inhabited by two groups of different nationality and religious backgrounds, Swedish Lutherans and German Catholics.   Among the Lutherans in the age class 40 to 45 years are 40 females and 44 males, among the Catholics 62 females and 58 males, all married to someone included in the enumeration.   The records show 18 mixed marriages, 11 between Lutheran males and Catholic females, and 7 between Catholic males and Lutheran females.   How does this observation compare with the number of mixed marriages that would be expected if there were no prejudice for or against them in the community? We set up the totals of Table 43.   By the definition on page 145, the probability of a marriage occurring in row (1) is $_1n/N = \frac{62}{102}$, and of a marriage occurring in col. (1) is $n_1/N = \frac{44}{102}$.   By

TABLE 43.—FOURFOLD TABLE FOR DETERMINING PROBABILITY OF MIXED MARRIAGES

Females	Males		
	Lutheran (1)	Catholic (2)	Total (3)
Catholic (1)......................	$_1f_1 = 26.7$	$_1f_2 = 35.3$	$62 = {}_1n$
Lutheran (2)......................	$_2f_1 = 17.3$	$_2f_2 = 22.7$	$40 = {}_2n$
Total (3)......................	$n_1 = 44$	$n_2 = 58$	$102 = N$

Theorem 3, the probability of a marriage occurring in both row (1) and column (1) is $(_1n/N)(n_1/N) = (\frac{62}{102})(\frac{44}{102})$; therefore the expected number of Catholic women marrying Lutheran men is $(_1n/N)(n_1/N)(N) = {}_1nn_1/N = 62(44)/102 = 26.7$.   This expected frequency is entered in the proper cell in the table. Similarly, the expected frequency in the cell common to row (2) and column (2) is $40(58)/102 = 22.7$.   Thus the total number of expected mixed marriages is $26.7 + 22.7 = 49.4$, or approximately 49; whereas, the observed number is 18, only 36 per cent of the expected number.   Evidently, there are obstacles in the way of marriages between the Swedish Lutherans and the German Catholics in this community.

This conclusion may be more fully established by applying the Chi-square (χ^2) method to Table 44. This method is designed to test the hypothesis that the differences between a set of observed and expected frequencies may be due solely to chance.

To obtain χ^2, we subtract each expected frequency (f_t) from the corresponding observed frequency (f_0), divide the squared difference by the expected frequency, and sum these ratios. The calculations are shown in Table 44.

TABLE 44.—CHI-SQUARE (χ^2) TEST

| Females | Males | Marriages | | $f_o - f_t$ | $(f_o - f_t)^2$ | $\dfrac{(f_o - f_t)^2}{f_t}$ |
		Observed (f_o)	Theoretical (f_t)*			
Catholic.....	Lutheran	11	26.7	−15.7	246.5	9.23
Catholic.....	Catholic	51	35.3	+15.7	246.5	6.98
Lutheran	Lutheran	33	17.3	+15.7	246.5	14.25
Lutheran	Catholic	7	22.7	−15.7	246.5	10.86
				0.0		41.32 = χ^2

* If any theoretical cell frequency is less than five, a correction is needed. See Paul Rider, *An Introduction to Modern Statistical Methods*, pp. 112–113, John Wiley & Sons, Inc., New York, 1939.

It was seen above that the expected frequencies used in Table 44 were calculated from the row and column totals of the observed frequencies in Table 43. This means that the observed and expected frequencies in the cells of Table 44 were to a certain extent *made* to agree. Evidently, this forced agreement should be allowed for in testing the amount of difference between the two sets of frequencies. In any 2 × 2 table, like Table 43, it is clear that if the row totals, the column totals, and *one* observed cell frequency are given, the other three cell frequencies are at once determined. Therefore, only one cell frequency is *free* of the influence of the marginal totals, so that a 2 × 2 table is said to have one *degree of freedom*.[1] If now the value of χ^2 obtained is referred to a table of χ^2, such as Appendix Table 2, that takes account of degrees of freedom, the spurious resem-

[1] The degrees of freedom for any contingency table are $(c - 1)(r - 1)$, where c is the number of columns and r is the number of rows. See A. E. Treloar, *Elements of Statistical Reasoning*, pp. 215 and 229, John Wiley &

blance between the observed and expected frequencies to which we objected above is corrected for.

Entering Appendix Table 2 with one degree of freedom, then, we find that a χ^2 as large as 6.635 could occur by chance once in 100 times, the theory involved here being similar to that described in the latter part of Sec. 4, below. Our χ^2 is 41.32, which is much larger, and would occur by chance less often than once in 100 times. Since it is customary to reject chance as the explanation of an event that can happen by chance no oftener than five times in 100, we conclude that the frequency of mixed marriages in the community cited is reduced by sociological and perhaps economic forces.

The classic method of introducing the elementary notions of probability is to use the illustration of coin tossing. The event is the occurrence of a "head" or a "tail." We may toss one coin several times, several coins once, or several coins several times, as we wish. It is evident that the events are mutually exclusive, as specified in Theorem 2, above. We may also assume that all the coins tossed during the experiment are exactly alike in size, weight, shape, and balance, *i.e.*, in respect to every fixed or biased factor that affects the tendency of heads or tails to fall uppermost when the coin is tossed. In this way we meet the requirement that each event of a probability set shall be *equally likely*. Differently expressed, it is assumed that the probability, p, of throwing a head is the same for every penny at each throw, and that every penny at each throw is independent of every other penny, *i.e.*, there is no tendency for one penny to show heads or tails because another does or does not, as would happen if they were stuck together. Finally, of the two events that can occur, one, heads, we call a *success*, and the other, tails, we call a *failure*. Having specified these conditions, our first question is, What is the probability of throwing a head, or of getting a success, at any one toss of a penny? In other words, what is the value of p?

Since in a single toss of one penny there is only one way in which a success can occur and one way in which a failure can

---

Sons, Inc., New York, 1939. A contingency table is a table of frequencies divided according to two or more principles of classification, such as the table in Exercise 7 at the end of this chapter.

[1] A "probability set" is described by the denominator of formula (38) or (39).

occur, and we assume that the pennies are balanced so that these two events are equally likely, we have, in the notation introduced above, $m = m' = 1$. Hence, from formulas (38) and (39), $p = q$, and substituting p for q or q for p in formula (40), we find that $p = q = \frac{1}{2} = 0.5$.

Suppose that we throw 10 pennies, and want to know the probability of getting exactly eight of a kind, *i.e.*, eight heads *or* eight tails. If eight of 10 pennies show heads, then the other two must show tails, or vice versa. We just saw that if we throw one penny, the probability of getting a head in one throw is $p = .5$. By Theorem 3, above, the probability of eight successes occurring in one possible way is $p^8 = (.5)^8$, the probability of two failures occurring in one possible way is $q^2 = (.5)^2$, and the probability of these two events occurring together in one possible way is $p^8q^2 = (.5)^8(.5)^2$. But the eight heads may occur among the 10 pennies in several possible ways, so that by Theorem 2 the probability of occurrence in just one way should be summed as many times as there are possible ways, or, more briefly, multiplied by the number of possible ways. How many possible ways are there? This is equivalent to asking, In how many ways may we get eight heads from 10 pennies, or, how many possible combinations are there of 10 $(= n)$ things taken eight $(= r)$ at a time? To answer this, we already have formula (37) above, which for our problem gives[1]

$$_{10}C_8 = \frac{10!}{2!8!} = \frac{10(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(8)(7)(6)(5)(4)(3)(2)(1)(2)(1)} = 45$$

Hence the probability, P, of getting exactly eight heads in a single throw of 10 pennies is

$$P = {_nC_r}p^rq^{n-r}, \tag{41}$$

or

$$P = 45(.5)^8(.5)^2 = 45(.5)^{10}.$$

Using logarithms,[2] we find

$$\log (.5)^{10} = 10 \log .5 = 10(9.69897 - 10) = 96.98970 - 100.$$

[1] See Appendix Table 3. For extensive table of factorials or their logarithms, see T. C. Fry, *Probability and Its Engineering Uses*, pp. 427–438, D. Van Nostrand Company, Inc., New York, 1928. A briefer table is given in *Mathematical Tables from Handbook of Chemistry and Physics*, 5th ed., p. 180, Chemical Rubber Publishing Company, Cleveland.

[2] See Appendix Table 7 and accompanying Foreword.

The antilogarithm of this is .0009766. Hence

$$P = 45(.0009766) = .044.$$

That is, in 44 out of 1,000 trials we would expect by the laws of chance to get exactly eight heads in a toss of 10 pennies. Similarly, by Theorem 2, the probability of getting either eight heads or eight tails is $2 \times 0.044 = 0.088$. This last is the probability that answers our question. Any similar question can be readily answered by substituting in formula (41), above.

3. The Binomial Distribution.—We often want to know the probability of getting as many as or more than a specified number of successes or failures. From what has been said, it will be seen that the probability of getting no successes at all in a toss of n pennes is q^n, of getting one success is $_nC_1pq^{n-1}$, of getting two successes is $_nC_2p^2q^{n-2}$, and so on, and, finally, the probability of getting all successes is p^n. Since these combinations of events exhaust the possibilities, some one of them is certain to occur at any toss of n pennies. In other words, the probability of one or another of them occurring is unity, or one. By Theorem 2, we may therefore write the equation

$$q^n + {}_nC_1pq^{n-1} + {}_nC_2p^2q^{n-2} + {}_nC_3p^3q^{n-3} + \cdots + {}_nC_rp^rq^{n-r}$$
$$+ \cdots + p^n = 1. \quad (42)$$

But by formula (40), $p + q = 1$, and hence $(p + q)^n = 1$. It therefore appears that by substitution

$$(q + p)^n = q^n + {}_nC_1pq^{n-1} + {}_nC_2p^2q^{n-2} + \cdots + {}_nC_rp^rq^{n-r}$$
$$+ \cdots + p^n. \quad (43)$$

If $p = q = \frac{1}{2}$, the formula simplifies to

$$(\tfrac{1}{2} + \tfrac{1}{2})^n = (\tfrac{1}{2})^n(1 + {}_nC_1 + {}_nC_2 + \cdots + {}_nC_{n-1} + 1). \quad (44)$$

This is the familiar binomial expansion of algebra, which is now seen to be an expression of the operation of the laws of chance![1]

---

[1] In algebra, the binomial formula is usually written:

$$(q + p)^n = q^n + \frac{n}{1}q^{n-1}p + \frac{n(n-1)}{1 \cdot 2}q^{n-2}p^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}q^{n-3}p^3$$
$$+ \cdots + p^n;$$

and it is pointed out that the exponent of q decreases by 1, while the exponent of p increases by 1, each term; and that the coefficient of any term, if multiplied by the exponent of q and divided by the number of the term, gives the coefficient of the next term.

To discover the probability of getting, say, eight or more heads in a single toss of 10 pennies, therefore, we need only apply the binomial. The probability of getting eight or more heads means, specifically, the probability of getting eight, nine, or 10 heads; and by Theorem 2, this is equal to the sum of the probabilities of the three separate events. By formula (41), which is the general term of the binomial, the probability of eight heads is $_{10}C_8 p^8 q^2$, of nine heads is $_{10}C_9 p^9 q$, and of 10 heads is p^{10}. Summing these,

$$P = {_{10}C_8 p^8 q^2} + {_{10}C_9 p^9 q} + p^{10} = 45(.5)^8(.5)^2 + 10(.5)^9(.5)$$
$$+ (.5)^{10} = (.5)^{10}(45 + 10 + 1) = .0440 + .0098 + .0010$$
$$= 0.055.$$

Accordingly, in 1,000 throws of 10 pennies, we may expect to get eight, nine, or 10 heads about 55 times. And the probability of getting eight or more heads *or* eight or more tails is, of course, $2 \times .055 = .11$, or 11 times in 100 throws. Notice that this is merely the most probable number and will vary from one set of 100 throws of 10 pennies each to another. But in a very large number of throws the average proportion should come rather close to 11 per 100 throws of 10 pennies each.

Suppose, again, we throw 10 pennies 150 times. In how many of these trials may we expect to get exactly eight of a kind? Since we have found this probability to be 0.088, we may expect this event in the proportion of about nine times in 100 trials, in the long run. If N represents the number of trials, and S the number of trials in which the specified event may be expected to happen, the formula is approximately

$$S = PN. \tag{45}$$

Substituting $P = .088$ and $N = 150$ in this formula, we find $S = .088(150) = 13.2$. That is, in 150 tosses of 10 pennies each, about 13 is the most probable number of tosses that will show exactly eight heads or eight tails.

Similarly, if it is wanted to know the frequency with which each possible number of successes, from 0 to n, may be expected to occur by chance in N trials of n events each, each term of the binomial expansion in formula (42) or (43) is simply multiplied by N:

$$q^n N + {_nC_1 pq^{n-1}N} + {_nC_2 p^2 q^{n-2}N} + \cdots + {_nC_r p^r q^{n-r}N}$$
$$+ \cdots + p^n N = N. \tag{46}$$

Thus, if we throw 10 pennies 1,000 times, we have

$1,000(.5 + .5)^{10} = .00098(1,000) + .00977(1,000) +$
$.04395(1,000) + .11719(1,000) + .20508(1,000) + .24609(1,000)$
$+ .20508(1,000) + .11719(1,000) + .04395(1,000)$
$+ .00977(1,000) + .00098(1,000),$

or

$.98 + 9.77 + 43.95 + 117.19 + 205.08 + 246.09 + 205.08$
$+ 117.19 + 43.95 + .98 = 1,000.$

This is really a binomial frequency distribution[1] and is so arranged in Table 45. From this table, we see that out of 1,000 tosses of 10 pennies each, we would expect no heads in only about one toss, one head in something like 10 tosses, two heads in approximately 44 tosses, and so on.

TABLE 45.—FREQUENCY DISTRIBUTION OF 1,000 TOSSES OF 10 PENNIES

Number of Heads (X)	Number of Tosses (f)
0	.98
1	9.77
2	43.95
3	117.19
4	205.08
5	246.09
6	205.08
7	117.19
8	43.95
9	9.77
10	.98
Total...............	1,000.00

Let us now pass from the theoretical case of penny tossing to some problem that might arise in social research. For example, the proportion of males in the urban population of Wisconsin in 1930 was $p_1 = 0.4974$; in the rural nonfarm population, $p_2 = 0.5118$; and in the rural farm population, $p_3 = 0.5435$. If we regard the three populations—urban, rural nonfarm, and rural farm—as ranked in the order of urbanness, and if we subtract the proportion of males in the less urban from that in the more urban of each of the three possible pairings of these populations, we get

[1] See Chap. V.

$$p_1 - p_2 = .4974 - .5118 = -.0144.$$
$$p_1 - p_3 = .4974 - .5435 = -.0461.$$
$$p_2 - p_3 = .5118 - .5435 = -.0317.$$

We notice that all three of the signs are negative. In the case of another Middle Western state taken at random, the same result was found. Does this mean that the proportion of males is really greater in the more rural populations, or may the negative signs in the two states be just a trick of chance? By formula (36), we see that there are $_3P_3 = \dfrac{3!}{0!} = 6$ possible orders of relative magnitude that p_1, p_2, and p_3 can take (*e.g.*, $p_1 < p_3 < p_2$; $p_2 < p_1 < p_3$; etc.), if we assume that they are never equal (*i.e.*, $p_1 \neq p_2 \neq p_3$). Since the order observed, $p_1 < p_2 < p_3$, is only one of the six, the probability that it will occur in one random trial (or state) is $\frac{1}{6}$. Hence the probability of getting only negative signs in both states is

$$_2C_2(\tfrac{1}{6})^2(\tfrac{5}{6})^0 = (\tfrac{1}{6})^2 = \tfrac{1}{36} = .028$$

by formula (41).

Statisticians usually insist on odds of at least 5 in 100, or 0.05, before they will risk the assumption that a result is not due to chance. By this standard, we eliminate chance in the present case, and are entitled to conclude that the proportion of males in the three populations is related to the degree of urbanness in those populations.

In many situations similar to this, the binomial theorem enables us to determine the probability that repeated events may occur by chance alone, and to note whether or not the probability is so small that we may reject the hypothesis that chance is responsible.

It is important to ask what is meant by chance in the preceding illustration. If we regard the census figures for the three populations as representing three complete universes, there is no question of chance at all. Any differences noted in the proportions of males, however small they may be, are real differences between the universes, and that is the end of the matter. But if we think of the proportion of males in each of our three populations as determined by a separate set of causes acting to produce sample results, and if we want to know whether or not these

three sets of forces differ in any real way from one another, the problem of chance at once enters.   By chance we mean a great number of small, unknown factors acting in many directions, as contrasted with large (biased)[1] factors, usually known or knowable, acting constantly in the same direction.   If the biased factors affecting the proportion of males differ from one of the three populations to another—*e.g.*, more females than males migrate from rural to urban areas—the observed proportions of males will differ to a greater extent than can be accounted for by the action of small random forces.   If the biased factors that produce the proportion of males in each of the three populations are essentially the same, however, any variation in the proportion of males from one population to another must be due to chance factors alone.   It is usually good research method to seek to eliminate chance as a possible cause of differences before undertaking to discover what factors are responsible.

If we already know from independent evidence, however, that important factors influencing the proportion of males varied between the three populations—*e.g.*, the two sexes migrated unequally from the more-rural to the less-rural areas—there would be no point in testing the hypothesis that the differences were due to chance, except perhaps to confirm the a priori knowledge.   When such a test fails to eliminate chance, it often means only that a larger sample is needed.   It may sometimes be advisable to investigate carefully the biased factors in the situations under comparison, even when chance has not been eliminated as a possible cause of the differences observed between them.

A binomial distribution, such as that of Table 45, is like other distributions in having a mean, a standard deviation, and other statistical constants by which it may be described.   The formulas for the mean and the standard deviation are

$$M_B = np, \qquad\qquad (47)$$
$$\sigma_B = \sqrt{npq}, \qquad\qquad (48)[2]$$

where the symbols have the same meanings as above.

For the distribution of Table 45, $M_B = 10(.5) = 5$ heads, and $\sigma_B = \sqrt{10(.5)(.5)} = 1.58$ heads.

[1] See also third paragraph on p. 149, above.

[2] For a derivation of these formulas see, for example, C. H. Richardson, *An Introduction to Statistical Analysis*, pp. 228–230, Harcourt, Brace and Company, Inc., New York, 1934.   The subscript, *B*, means *binomial*.

It is not necessary in chance situations that p and q should be equal. Thus, the probability of throwing an ace in a single toss of a die is $p = \frac{1}{6}$, and the probability of not throwing an ace is $q = \frac{5}{6}$. If 15 throws are to be made, the binomial is $(\frac{1}{6} + \frac{5}{6})^{15}$, and this can be expanded and utilized just as was done above for $p = q = \frac{1}{2}$. When $p = q$, the binomial is symmetrical in shape, when $p \gtrless q$[1], it is asymmetrical or skewed.

4. The Normal Distribution.—Graphs of the binomials $32(\frac{1}{2} + \frac{1}{2})^5$ and $1,024(\frac{1}{2} + \frac{1}{2})^{10}$ are shown in Fig. 38. Notice that
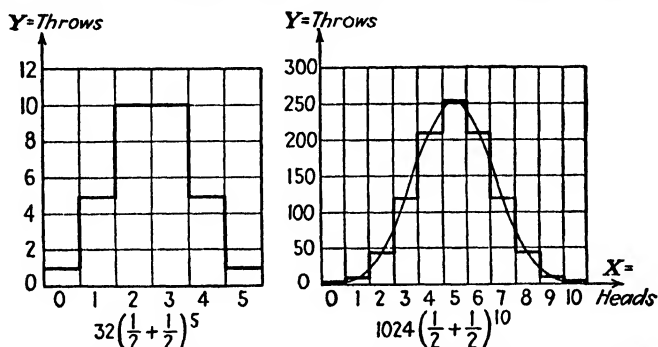
FIG. 38.—Histograms of two binomials, $N(\frac{1}{2} + \frac{1}{2})^n$, as n increases.
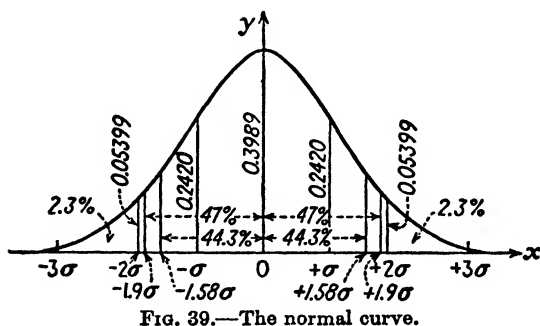
FIG. 39.—The normal curve.

they take the form of histograms rather than of smooth curves, because successes are counted only in whole numbers, yielding a discrete or discontinuous series. However, if the length of the scale is kept constant, as in the figure, the graph of the binomial $1,024(\frac{1}{2} + \frac{1}{2})^{10}$ is seen to be less broken in outline than is that of the binomial $32(\frac{1}{2} + \frac{1}{2})^5$. As n increases, the graph approaches closer and closer to a smooth curve in appearance. If now n is indefi-

[1] $\gtrless$ means greater or less than.

nitely increased, giving the binomial $N(\frac{1}{2} + \frac{1}{2})^{\infty}$, it is evident that the intervals of the graph become smaller and smaller, until in effect the outline merges into that of a smooth curve. The resulting curve is the most important type of distribution in statistical theory, and is known variously as the *normal curve*, the *Gaussian curve*, the *curve of error*, or the *curve of probabilities*. Unlike the binomial distribution, it represents a continuous variable, which can take any value whatever, on the X scale. A graph of the normal curve is shown in Fig. 39. It may be thought of as enclosing a continuous surface, cut from a piece of thin sheet metal. Its equation is usually written

$$y = \frac{N}{\sigma_x \sqrt{2\pi}} e^{\frac{-x^2}{2\sigma_x^2}},$$ (49)

where $x = X - M$, or a mean deviate of X,

N = total frequency of the distribution,

π = 3.1416, so that $\sqrt{2\pi}$ = 2.5066,

e = 2.7183, the base of natural logarithms.

If the area of the curve is taken as unity, equation (49) becomes

$$y = \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-x^2}{2\sigma_x^2}}$$ (50)

As an aid to understanding the curve represented by equation (50), let us analyze its equation. We shall begin by letting $\frac{x}{\sigma_x} = t$, so that equation (50) becomes

$$y = \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-t^2}{2}}$$ (51)

In the calculation of tables of normal ordinates, it is also convenient to let $\sigma_x = 1$, giving

$$y = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$$ (52)

But, as seen above, π is a mathematical constant with the value 3.1416, so that $\sqrt{2\pi}$ = 2.5066, and $\frac{1}{2.5066}$ = .3989. Equation (52) may therefore be written

$$y = .3989 e^{\frac{-t^2}{2}}$$ (53)

In Fig. 39, at the mean of the X's on the X axis, $x = 0$. The height of the ordinate at any point is the value of y found from equation (53) by substituting the appropriate value of t. At $x = 0$, $t = x/\sigma_x = 0/\sigma_x = 0$, and

$$y = .3989e^{\frac{-(0)^2}{2}}$$
$$y = .3989e^0.$$

But any number raised to the zero power is 1*, so that

$$y = .3989(1) = .3989.$$

In other words, at the mean of the X's, the height of the ordinate, y, is .3989, for any normal curve of unit area and unit standard deviation. This is plotted in Fig. 39.

Next, for the same case, let $t = x/\sigma_x = +2$. Then, by formula (53),

$$y = .3989e^{\frac{-(2)^2}{2}},$$
$$y = .3989e^{-\frac{4}{2}},$$
$$y = .3989e^{-2}.$$

But

$$e^{-2} = \frac{1}{e^2},$$

so that

$$y = \frac{.3989}{e^2}.$$

It has also been seen that e, like π, is a mathematical constant, having the value 2.7183, so that $e^2 = 7.38906$. Hence, at

$$t = \frac{x}{\sigma_x} = +2,$$
$$y = \frac{.3989}{7.38906} = .05399.$$

This value is also plotted in Fig. 39. Notice that at $\frac{x}{\sigma_x} = -2$, the value of y is the same as at $\frac{x}{\sigma_x} = +2$, for in formula (53) evidently $e^{-(2)^2}$ is the same as $e^{-(-2)^2}$. The normal curve is thus symmetrical, *i.e.*, of the same shape, on each side of the mean. From this it follows that the mean and the median coincide.

* See any text in elementary algebra.

The student is asked to check the values of y found above at $x/\sigma_x = 0$ and at $x/\sigma_x = \pm 2$ against those printed in Appendix Table 1.   All the values in that table are calculated in this way, and may be used to complete the construction of Fig. 39.   Thus, the height of the ordinates at $\pm 1\sigma$, read from the table, is .2420, and is so scaled in the figure.   After several ordinates have been drawn, they are connected by a smooth line, to form the curve shown.

The tallest ordinate of the normal curve occurs at the mean, hence the mean, median, and mode all coincide.   This appears from the fact that when $x = 0$, $y = .3989$; whereas, when $x \gtrless 0$, $y = .3989/e^{\frac{t^2}{2}}$.   The latter term is always smaller than the former, since all positive powers of e are greater than $1(e^0 = 1)$.

Another characteristic of the normal curve is that it is asymptotic to the X axis, meaning that the curve constantly approaches but never touches the X axis as it extends indefinitely in both directions from the mean.

Table 46 shows a hypothetical normal distribution with perfectly symmetrical frequencies.   The actual frequencies of normal tables may depart in various degrees from this symmetrical pattern, because of sampling errors or the use of class intervals that do not place the mean of the series exactly at the center of the distribution.

TABLE 46.—NORMAL DISTRIBUTION OF SCORES ON AN ARMY ATTITUDES TEST
(HYPOTHETICAL DATA)

Scores (X)	Men (f)
0– 4.9	5
5– 9.9	17
10–14.9	44
15–19.9	92
20–24.9	150
25–29.9	191
30–34.9	191
35–39.9	150
40–44.9	92
45–49.9	44
50–54.9	17
55–59.9	5
Total......................	998

With the help of the integral calculus, it is possible to find the proportion of the area under any part of the normal curve, *i.e.*, between the ordinates erected at any two points on the x scale. This has been done for the areas between the ordinate at the mean and ordinates erected at intervals of $.01\sigma$ along the x-axis. The results are shown in Appendix Table 1, in the column headed "Area." Thus the area under the curve between the ordinate at the mean and the ordinate at 1σ is seen to be 0.34, or 34 per cent (roughly one-third) of the total area under the curve. In Chap. VI we saw that the area under a frequency histogram, where the width of the interval is taken as one unit, is equal to the total frequency of the distribution. The same principle holds for the normal curve.

Since the normal curve represents the distribution of frequencies in any normal universe, the proportion of the area between the ordinate at the mean and the ordinates at, say, $x = \pm 1\sigma$ represents the most probable proportion of the frequencies of any random sample drawn from such a universe that may be expected to fall between the values $\frac{x}{\sigma} = 0$ and $\frac{x}{\sigma} = \pm 1$. Differently expressed, the proportion of the area between the ordinate at the mean and the ordinates at $x = \pm 1\sigma$ is the *probability* that a random sample value of X will fall between M_x and $\pm 1\sigma$. We see from Appendix Table 1 that this probability is twice 0.34, which is approximately 0.68, or 68 per cent. It should now be clear why in a normal distribution the odds are about two to one that a random value of X will be within a range of one standard deviation on each side of the mean value of X. Also, inasmuch as a value of X falls outside the range of $M_x \pm 2\sigma$ by chance only $1.00 - (2 \times 0.477) = 0.046$, or about one time in 20, we shall be fairly safe if we attribute those values that do so to something else than chance. In other words, we shall arbitrarily regard all such extreme values as *significant*.

Reading again from Appendix Table 1, it is seen that approximately 25 per cent of the area of the normal curve lies between the mean and an ordinate at $x = 0.67\sigma$. That is, one-half of the area of the curve is included between an ordinate at -0.67σ and an ordinate at $+0.67\sigma$. From a finer table, the figure is found more exactly to be 0.6745σ. The distance 0.6745σ from the mean along the x-axis of the normal curve is commonly called the

probable error (*P.E.*), and is often used instead of the standard deviation, σ, or *standard error*, as it is called in sampling theory (see Chap. XII).

The relationships of the preceding paragraphs do not hold, however, for skewed distributions. This may be seen from Fig. 41. By comparing the rectangles in the areas $M - 1\sigma$ and $M + 1\sigma$, it is clear that in this case a much larger proportion of the area of the curve is contained between $M - 1\sigma$ than between $M + 1\sigma$, so that the standard deviation has no constant relation to the area or frequency. For this reason, the standard deviation has a variable meaning when applied to asymmetrical distributions, and should be cautiously interpreted in such cases.
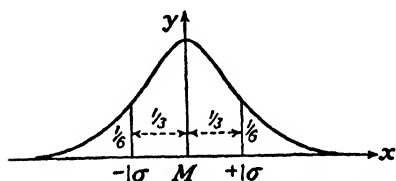
Fig. 40.—Relation between standard deviation and area under normal curve.
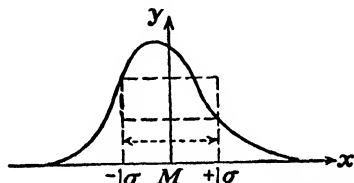
Fig. 41.—Relation between standard deviation and area under skewed curve.

In a normal distribution, $A.D. = .80\sigma$, so that the distance $M \pm A.D.$ on the scale includes about 58 per cent of the frequencies (see Appendix Table 1).

When n is large, the labor of expanding the binomial becomes excessive. Under these conditions, if the value of np or nq is not too small, say 5 or more, the binomial so closely approximates the normal curve that the latter may be used in its stead for purposes of estimation, and the desired probabilities simply read from Appendix Table 1.

Consider again the probability of getting eight or more heads or tails in a toss of only 10 pennies. In Fig. 39 we erect a perpendicular at the point

$$\frac{x}{\sigma} = \frac{X - \bar{X}}{\sigma_B} = \frac{X - np}{\sqrt{npq}} = \frac{8 - 10(.5)}{\sqrt{10(.5)(.5)}} = 1.9,$$

and find the area between this ordinate and the ordinate at the mean. Entering Appendix Table 1 with $x/\sigma = 1.90$, we see that the area desired is 0.4713 of the area of the whole curve. Subtracting 0.4713 from 0.5000, the area of half the curve, we get 0.0287 as the area to the right of the ordinate at $x/\sigma = +1.9$.

Since the normal curve is assumed to represent the results of all possible tosses of 10 pennies, the area to the right of $x/\sigma = +1.9$ shows the proportion of tosses that in the long run may be expected to give eight or more heads. This proportion is the *probability* of getting eight or more heads in one toss of 10 pennies, so the probability of getting eight or more heads or eight or more tails is twice this, or $P = 2(0.0287) = 0.0574$. The true value of P as found above from the binomial expansion is $P = 0.1094$. The agreement is thus seen to be none too good when n is as small as 10. If n is increased to 15, however, $np = 7.5$, and we find more agreement. The probability of getting say 12 or more heads or tails is 0.0204 according to the normal curve, and 0.0176 according to the binomial,[1] the error being only 0.0028. For larger values of n, the two estimates may for most purposes be accepted as equivalent.

The approximate probability of getting *exactly* eight heads or eight tails in a toss of $n = 10$ pennies is the height of the ordinate of the normal curve at the point $X = 8$, expressed in standard deviation units. This is because the number 8 is represented on the X scale by a point rather than by a distance, and on this point can be erected only a straight line, or ordinate, which theoretically has no width and hence no area. We now need $\frac{y}{\sigma_x}$ at $X = 8$, *i.e.*, at $\frac{x}{\sigma} = \dfrac{8 - 10(0.5)}{\sqrt{10(0.5)(0.5)}} = 1.9$. From Appendix Table 1 we find $y = 0.0656$, so that $\dfrac{y}{\sigma_x} = \dfrac{0.0656}{\sqrt{10(0.5)(0.5)}} = 0.0415$, and $2 \times 0.0415 = 0.083$ is the probability desired. The correct probability already found by the binomial is 0.0879.

If we choose to consider the normal curve merely as a device for approximating the probabilities of the binomial, rather than as a continuous mathematical distribution, it becomes possible to take certain liberties with it that will improve its accuracy for the purpose. For example, to determine the probability of throwing eight or more heads or tails in a toss of 10 pennies, we may allow the value $X = 8$ to occupy the area under the normal curve between the X values 7.5 and 8.5, and regard the area to the right of 7.5 as representing the probability of throwing eight or more heads. We may then erect a perpendicular in Fig. 39

[1] $_{15}C_{13}p^{13}q^3 + {}_{15}C_{13}p^{13}q^2 + {}_{15}C_{14}p^{14}q + p^{15} = (\tfrac{1}{2})^{15}(455 + 105 + 15 + 1)$
$= 0.01758$.

at the point

$$\frac{x}{\sigma} = \frac{X - np}{\sqrt{npq}} = \frac{7.5 - 10(0.5)}{\sqrt{10(0.5)(0.5)}} = 1.58,$$

and find the area between this ordinate and the ordinate at the mean to be 0.4429 (Appendix Table 1). The area to the right of the ordinate at 1.58σ is, therefore, $0.5000 - 0.4429 = 0.0571$, which is the probability of throwing eight or more heads. The probability of throwing eight or more heads or eight or more tails is $2 \times 0.0571 = 0.1142$. This result is much closer to the correct binomial probability of 0.1094 than was that obtained above in the orthodox way. Indeed, the accuracy of the normal curve in approximating the binomial has now been made quite satisfactory even for $n = 10$.

It is also possible to use a similar manipulation in estimating the probability of throwing exactly eight heads or eight tails in a toss of 10 pennies. We find from Appendix Table 1 the area under the curve included between an ordinate at $X = 7.5$ and an ordinate at $X = 8.5$. The table gives 0.4864 as the area between the mean ordinate and the ordinate at $X = 8.5$ (*i.e.*, at $x = \dfrac{8.5 - 10(0.5)}{\sqrt{10(0.5)(0.5)}} \sigma = 2.21\sigma$), and 0.4429 as the area between the mean ordinate and the ordinate at $X = 7.5$ (*i.e.*, at

$$x = \frac{7.5 - 10(0.5)}{\sqrt{10(0.5)(0.5)}} \sigma = 1.58\sigma).$$

Consequently, the area between the ordinate at $X = 7.5$ and the ordinate at $X = 8.5$ is

$$0.4864 - 0.4429 = 0.0435.$$

This is the probability of throwing exactly eight heads in a toss of 10 pennies; so the probability of throwing exactly eight heads or exactly eight tails is $2 \times 0.0435 = 0.0870$. The error from the binomial (0.0879) in this case is negligible.

It should be noted that special modifications like those above in the use of the normal curve are usually worth while only when np is small, say $np < 6$.

5. Skewness and Kurtosis.—The frequency distributions with which social scientists have to deal usually depart considerably from the normal form. Such a distribution is shown

in Table 47 and in Fig. 42. It is readily seen to extend farther in the positive direction from the mean than in the negative
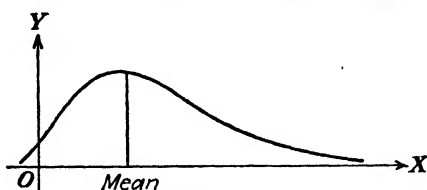
FIG. 42.—Skewed distribution.

direction, and so is said to be *positively skewed*. If there is occasion to measure the amount of the skewness, an index is

TABLE 47.—RELATIVE NUMBERS OF DIVORCED COUPLES BY YEARS MARRIED

Years married (X)	Divorced couples (f)	Accumulated frequency
0–0.9	15	15
1.0–1.9	72	87
2.0–2.9	60	147
3.0–3.9	43	190
4.0–4.9	21	211
5.0–5.9	17	228
6.0–6.9	9	237
7.0–7.9	8	245
8.0–8.9	5	250
9.0–9.9	2	252
Total.........	252	

provided by formula (54):

$$Sk = \frac{3(M - Md)}{\sigma}. \tag{54}[1]$$

[1] We saw in Chap. VII that the value of the mean, M, is influenced by extreme values, and hence by skewness, but that the value of the mode, Mo, is not affected. In the present chapter it was learned that in a normal distribution the mean, mode, and median all have the same value. These facts suggest as an approximate measure of absolute skewness, Sk, the difference

$$Sk = M - Mo. \tag{55}$$

To change this to generalized units, we may write

$$Sk = \frac{M - Mo}{\sigma}. \tag{56}$$

Because the value of the mode can seldom be accurately determined, how-

The values of Sk by this formula vary between ± 3, but values larger than ± 1 do not often occur. If there is no skewness, $Sk = 0$.

A more useful measure of skewness for some purposes is g_1, which for large samples is approximately

$$g_1 = \frac{\nu_3}{\sigma^3}. \tag{58}[1]$$

ν_3 is the *third moment* about the mean of the distribution, defined by the equation $\nu_3 = \Sigma f x^3 / N$, where x is a mean deviate as usual.

For a normal distribution $g_1 = 0$. For other values of g_1 the sign indicates the direction of the skewness. Values of g_1 as great as ± 2 mean decided skewness.

A frequency distribution may also depart from the normal in height or "peakedness." This is called *kurtosis*. If the observed distribution is flatter than the normal, it is said to be *platykurtic;* if more peaked, *leptokurtic;* if neither, *mesokurtic.* Kurtosis may be measured by g_2. For large samples, an approximate formula is

$$g_2 = \frac{\nu_4}{\sigma^4} - 3 \tag{59}$$

ν_4 is the *fourth moment*, $\Sigma f x^4 / N$, of the distribution; and σ^4 is the *second moment*, $\nu_2 = \sigma^2 = \Sigma f x^2 / N$, squared.

g_2 also is zero for a normal distribution. A positive value of g_2 indicates that the observed distribution is more peaked than the normal, and a negative value indicates that it is flatter.

---

ever, it is considered preferable to replace it by its equivalent in terms of the median, Md. In any moderatey skewed distribution, the median falls about two-thirds of the distance from the mode to the mean (see Chap. VII, Fig. 31). We therefore have

$$Mo = M - 3(M - Md). \tag{57}$$

Substituting this value of the mode in formula (56),

$$Sk = \frac{M - [M - 3(M - Md)]}{\sigma},$$

$$Sk = \frac{3(M - Md)}{\sigma}. \tag{54}$$

[1] ν is the lower-case Greek letter nu.

Before formula (58) or (59) can conveniently be applied to a distribution like that of Table 47, some short-cut calculating formulas are needed:

$$\nu_2 = \sigma^2 = \frac{i^2}{N}\left[\Sigma fd^2 - \frac{(\Sigma fd)^2}{N}\right] \tag{60}$$

$$\nu_3 = \frac{i^3}{N}\left[\Sigma fd^3 - \frac{3}{N}\Sigma fd\,\Sigma fd^2 + \frac{2}{N^2}(\Sigma fd)^3\right] \tag{61}$$

$$\nu_4 = \frac{i^4}{N}\left[\Sigma fd^4 - \frac{4}{N}\Sigma fd^3\Sigma fd + \frac{6}{N^2}\Sigma fd^2(\Sigma fd)^2 - \frac{3}{N^3}(\Sigma fd)^4\right], \tag{62}$$

where i = width of class interval.

d = unit step deviation from an assumed mean.

$N = \Sigma f$.

Notice that formulas (61) and (62) are merely extensions of the familiar short method of finding a standard deviation by the use of an assumed mean and unit step intervals. This appears clearly in Table 48, below.

Let us now measure the skewness and kurtosis of the distribution shown in Table 47, by comparing it with the normal curve. We set up the computing table:

TABLE 48.—COMPUTING TABLE FOR MOMENTS: DATA OF TABLE 47

Years married	f	d	fd	fd^2	fd^3	fd^4
0–0.9	15	−2	− 30	60	− 120	240
1.0–1.9	72	−1	− 72	72	− 72	72
2.0–2.9	60	0	0	0	0	0
3.0–3.9	43	+1	+ 43	43	43	43
4.0–4.9	21	+2	+ 42	84	168	336
5.0–5.9	17	+3	+ 51	153	459	1,377
6.0–6.9	9	+4	+ 36	144	576	2,304
7.0–6.9	8	+5	+ 40	200	1,000	5,000
8.0–8.9	5	+6	+ 30	180	1,080	6,480
9.0–9.9	2	+7	+ 14	98	686	4,802
Total..........	252	...	154	1,034	3,820	20,654

Recalling the short formula for the mean,

$$M = A + \frac{i\Sigma fd}{N},$$

where A is the assumed mean, we find for this table,

$$M = 2.5 + 1(\tfrac{154}{252}) = 3.11.$$

Substituting in the formula for the median,

$$Md = L + \left(\frac{N/2 - F}{f}\right) i,$$

where the symbols have the meanings explained in Chap. VII. We find

$$Md = 2.0 + \frac{126 - 87}{60} (1) = 2.65.$$

For the standard deviation, we have

$$\sigma = i \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2},$$

$$\sigma = 1 \sqrt{\tfrac{1034}{252} - (\tfrac{154}{252})^2},$$

$$\sigma = 1.93.$$

Hence, according to formula (54), we find the skewness to be

$$Sk = \frac{3(3.11 - 2.65)}{1.93} = 0.72.$$

This shows considerable skewness in the positive direction.

Let us next measure the amount of skewness in Table 48 by the use of formula (58). From formulas (60) and (61) we find

$$\sigma^2 = (1.93)^2 = 3.73,$$

$$\nu_3 = \frac{1}{252}\left[3820 - \frac{3}{252} (154)(1034) + \frac{2}{(252)^2} (154)^3 \right] = 8.09.$$

Substituting in formula (58),

$$g_1 = \frac{8.09}{(1.93)^3} = 1.13.$$

This result agrees with that obtained by formula (54), in showing positive skewness.

We shall now measure the degree of kurtosis, if any, exhibited by the distribution of Table 48, through the use of formula (59). We need only one new value, ν_4, which may be found by formula (62).

$$\nu_4 = \frac{1}{252}\left[20,654 - \frac{4}{252} (3,820)(154) + \frac{6}{(252)^2} (1,034)(154)^2 \right.$$

$$\left. - \frac{3}{(252)^3} (154)^4 \right],$$

$$\nu_4 = \frac{13,528}{252} = 53.7.$$

Substituting in formula (59),

$$g_2 = \frac{53.70}{(3.73)^2} - 3.00$$
$$g_2 = 3.86 - 3.00 = 0.86$$

The value of g_2 is positive, so we conclude that the observed distribution is leptokurtic, or more peaked than a normal curve.[1]

Even though a sample distribution is found by the above methods to differ from the normal, the question arises whether or not the difference is one that might be due merely to random errors of sampling. This point is dealt with in Chap. XIII.

Exercises

1. Twelve children are to be used in the experimental study of dominating and submissive types of behavior. Each child is to be grouped (*a*) with one other child, (*b*) with two other children. What is the total possible number of such experimental groups of each size?

2. Four villages, five cities, and five rural counties are to be grouped in all possible combinations of five. No distinction is made between areas of the same type, *i.e.*, one village is the equivalent of another village. What is the total number of combinations? Describe them.

3. The types of contact between families in a community are listed as: visit, church, lodge, school, business, and "other." But any or all of these contacts may appear together, as well as separately. How many combinations of all kinds are there between these several types of contact?

4. The educational levels of a sample of husbands and wives are recorded as college, high school, grades, and illiterate. What is the total number of possible permutations of husband-wife relationships in terms of these levels, and what are they?

5. How many marriages are possible between three pairs of brothers and sisters in our society?

---

[1] Another measure of kurtosis that is more commonly used than g_2 is β_2:

$$\beta_2 = \frac{\nu_4}{\sigma^4}. \tag{63}$$

For Table 48, above, $\beta_2 = 3.86$. Since in a normal distribution $\beta_2 = 3$, the observed distribution is again seen to be leptokurtic.

6. In an experiment with four pairs of subjects, each pair consists of a male and a female, closely "matched" in respect to certain sociological characteristics. They are to be given a test while seated around a table in such a way that the sexes alternate, and no members of a matched pair sit next to each other. In how many ways may this be done?

NOTE: The number of different permutations of n things taken n at time when arranged in a circle is given by the formula $(n - 1)!$

7. Gist and Clark give the following table:

DISTRIBUTION OF INTELLIGENCE SCORES OF 2,544 (KANSAS) RURAL HIGH-SCHOOL STUDENTS IN 1923, ACCORDING TO PRESENT RURAL AND URBAN CLASSIFICATION*

I.Q.	Urban	Rural	Total
Under 95...............................	378	832	1,210
95–104................................	326	472	798
105 and over..........................	260	276	536
Total...............................	964	1,580	2,544

* *American Journal of Sociology*, July, 1938, p. 43.

Compare the observed frequencies with those expected by chance alone, apply the χ^2 test, and comment on the results.

8. Classification of many cases shows that the probability of a marriage ending in divorce under certain conditions is 0.20. In a sample of 20 such marriages, what is the probability that there will be no divorce? What is the probability that there will be no more than two divorces? Compare the results from the binomial with those from the normal curve.

9. In Exercise 8, if many random samples of 20 marriages each were taken from the type of marriage referred to, (*a*) What *mean* number of marriages per sample would be expected to end in divorce? (*b*) What would be the standard deviation of the numbers of marriages ending in divorce found from many samples?

10. Calculate skewness and kurtosis for the distributions below:

FAILURES ON PAROLE IN 50 SUBSAMPLES OF FIVE PRISONERS EACH

Failures	Frequency
0	1
1	10
2	17
3	15
4	7
5	0
Total...................	50

FAMILIES BY SIZE

Persons	Frequencies
1	24
2	70
3	62
4	52
5	36
6	23
7	14
8	8
9	5
10	3
11	1
12 or more	1
Total	299

FAMILIES CLASSIFIED BY AGE OF MAN HEAD

Age, years	Frequency
Under 25	13
25–34	59
35–44	71
45–54	57
55–64	37
65–74	19
75 and over	6
Total	262

References

BATEN, W. D.: *Mathematical Statistics*, Chaps. IV, V, VI, and VII, John Wiley & Sons, Inc., New York, 1938.

CAMP, B. H.: *Mathematical Part of Elementary Statistics*, Part I, Chaps. II and V; Part II, Chaps. I, II, D. C. Heath & Company, Boston, 1931.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chap. X, pp. 333–337, Prentice-Hall, Inc., New York, 1939.

FINE, H. B.: *College Algebra*, pp. 393–407, Ginn and Company, Boston, 1904.

FRY, T. C.: *Probability and Its Engineering Uses*, Chaps. I, II, and III, D. Van Nostrand Company, Inc., New York, 1928.

HOLZINGER, K. J.: *Statistical Methods for Students in Education*, Chaps. XI–XII, Ginn and Company, Boston, 1928.

RICHARDSON, C. H.: *Introduction to Statistical Analysis*, Chaps. V, IX, and X, Harcourt, Brace and Company, Inc., New York, 1934.

SMITH, JAMES G.: *Elementary Statistics*, Part IV, Henry Holt and Company, Inc., New York, 1934.

TIPPETT, L. H. C.: *The Methods of Statistics*, 2d ed., Chaps. I, II, and IV, Williams and Norgate, Ltd., London, 1937.

TRELOAR, A. E.: *Elements of Statistical Reasoning*, Chaps. V, VI, XII, XV, and XVI, John Wiley & Sons, Inc., New York, 1939.

YULE, G. U., and M. G. KENDALL: *An Introduction to the Theory of Statistics*, Chaps. IX, X, and XXII, Charles Griffin & Company, Ltd., London, 1937.

CHAPTER X

GROSS RELATIONSHIP BETWEEN TWO FACTORS: SIMPLE LINEAR QUANTITATIVE CORRELATION

One of the most common purposes of social research is to discover whether or not there is any relationship between two factors, and to measure the amount of the relationship. For example, does the number of children in a family tend to decrease as the family income increases? If treated statistically, this kind of question is called a problem in *correlation*. As will be seen below, statistics is able to measure the amount of relationship (correlation) present in such cases, to provide an equation by which one of the factors can be predicted from a knowledge of the other, and to estimate the range of error in the predictions.

1. The Scatter Diagram: Ungrouped Data.—As an introduction to the method of simple linear correlation applied to ungrouped data, let us test the idea that the largest percentage increases of population in the United States between 1920 and 1930 occurred in regions where the density of population per square mile was least in 1920. We shall limit ourselves here to examining the amount of correlation in the nine census divisions. The necessary figures are given in Table 49.

TABLE 49.—PERCENTAGE OF POPULATION INCREASE, 1920–1930 (*Y*), IN RELATION TO POPULATION PER SQUARE MILE IN 1920 (*X*), BY GEOGRAPHIC DIVISIONS, UNITED STATES*

Division	X	Y
New England	119	10
Middle Atlantic	223	18
East North Central	88	18
West North Central	25	6
South Atlantic	52	13
East South Central	50	11
West South Central	24	19
Mountain	4	11
Pacific	18	47

* From Abstract of the Fifteenth Census of the United States, 1930, pp. 12–13.

We may make a preliminary judgment by rough methods as to whether or not any relationship is present between the X and Y series.   Taking the four largest values of X, we find the average of the four corresponding Y values to be 14.75.   For the four smallest values of X, the average Y value is 20.75.   In other words, as the X values decrease, the Y values tend to increase, on the average.   This suggests that there is some negative relationship between the two series.

A better way of prejudging correlation is by means of a *scatter diagram.*   The X and Y values are plotted on rectangular coordinate paper, as shown in Fig. 43.[1]   It is now seen that if
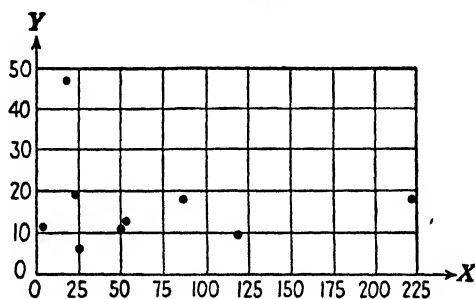
FIG. 43.—Scatter diagram for Table 49.

the point for the Pacific region is omitted, the remaining points show no discernible tendency either to rise or to fall across the table.   Any correlation present must, therefore, be due to a single case.   It would be misleading to say that between 1920 and 1930 there was a tendency for population in the United States to increase at a faster rate in thinly populated regions than in thickly populated regions, when as a matter of fact this was true in only one out of nine regions.   There is accordingly no point in going any further with this problem, unless we wish to try areas smaller than census divisions.

Consider a second problem.   Do the counties of Wisconsin that have high birth rates also tend to have high death rates? Waiving the objections that a county is not always a homo-

---

[1] For example, the first pair of values constitute a point with the *coordinates* (119, 10).   To plot this point in Fig. 43, after drawing the horizontal X axis and the Y axis perpendicular to it, we measure 119 units from the origin at 0 along the X axis, then up 10 Y units parallel to the Y axis, and there mark in the point.

geneous unit (*e.g.*, a county may be half urban and half rural), and that its population is often too small to yield reliable birth and death rates, let us compare the first 20 counties of the state, taken alphabetically, in 1935. The data are in Table 50.

TABLE 50.—BIRTH AND DEATH RATES BY COUNTIES IN WISCONSIN, 1935*

County	Birth rate (X)	Death rate (Y)	XY	X²	Y²
Adams...............	18.6	9.7	180.42	345.96	94.09
Ashland...............	22.2	12.0	266.40	492.84	144.00
Barron...............	18.4	10.4	191.36	338.56	108.16
Bayfield.............	12.5	8.3	103.75	156.25	68.89
Brown...............	22.1	11.6	256.36	488.41	134.56
Buffalo..............	17.5	6.9	120.75	306.25	47.61
Burnett.............	17.2	10.3	177.16	295.84	106.09
Calumet.............	15.7	6.8	106.76	246.49	46.24
Chippewa...........	20.5	12.1	248.05	420.25	146.41
Clark...............	17.3	7.4	128.02	299.29	54.76
Columbia...........	17.4	13.9	241.86	302.76	193.21
Crawford...........	22.5	10.1	227.25	506.25	102.01
Dane...............	17.1	13.8	235.98	292.41	190.44
Dodge..............	14.4	9.2	132.48	207.36	84.64
Door...............	20.8	9.8	203.84	432.64	96.04
Douglas............	16.2	12.2	197.64	262.44	148.84
Dunn...............	18.7	9.3	173.91	349.69	86.49
Eau Claire..........	22.0	12.2	268.40	484.00	148.84
Florence............	17.8	10.5	186.90	316.84	110.25
Fond du Lac.........	17.3	11.1	192.03	299.29	123.21
Total.............	366.2	207.6	3,839.32	6,843.82	2,234.78

* From *Report of the State Board of Health*, Wisconsin, 1934–1935, p. 210.

$$M_x = \frac{366.2}{20} = 18.31 \qquad M_y = \frac{207.6}{20} = 10.38$$

We shall apply the device of the scatter diagram to these figures. The results are shown in Fig. 44.

From Fig. 44, we notice first that the range taken by the points is limited, none falling below 12 or above 23 on the X scale, and none below 6 or above 14 on the Y scale. It is·a general precaution that as a rule any correlation found for a given set of data should not be assumed to exist outside the range of the data. A man may accept a wage of 50 cents an hour to work eight hours or perhaps even 12 hours without

resting, but it would be erroneous to suppose from this that he would continue to work an indefinite number of hours at that rate. After 12 or 14 hours, it would probably require more than 50 cents to induce him to work another hour. Thus the relationship between wages (X) and length of work period (Y) would not be the same beyond the range of 12 hours as within that range. Similarly, counties with birth rates much below 12 or above 23 might show death rates entirely out of line with what would be expected from the relationship found between birth and death rates in the counties included in the study.
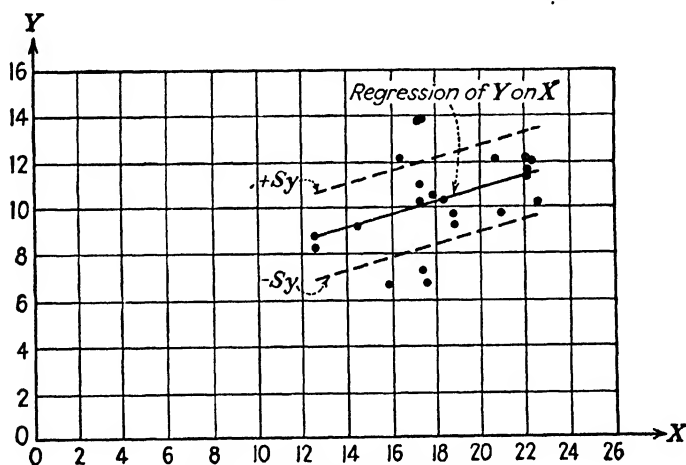
Fig. 44.—Scatter diagram for Table 50.

A second fact shown by Fig. 44 is that there is a *general* tendency for the points to rise in the positive direction along the X scale. That is, as the birth rates in the counties increase, the death rates tend to increase also. This indicates that there is some positive correlation between the two kinds of rates that seems worthy of further investigation. We would not expect a high correlation, however, because the dots show considerable scatter, instead of following one another in a continuous line or curve.

It should be pointed out that if the data in Fig. 44 had fallen instead of rising in the positive direction along the X scale, a negative relationship would have been indicated. That is, there would have been a tendency for the death rates to decline as the birth rates increased. A negative correlation, of course,

shows just as much relationship as a positive correlation of the same degree.

2. The Line of Regression: Ungrouped Data.—In simple correlation it is customary, whenever reasonable, to regard one of the factors, X, as an independent factor, and the other, Y, as a dependent factor. Thus, above, the birth rate is taken as the *independent* factor, X, and the death rate as the *dependent* factor, Y, because the birth rate is believed to influence the death rate, rather than vice versa.

Returning to Fig. 44, the next step in the attempt to measure the amount of correlation between the X and Y factors is to ask what is the *form* of the observed correlation. From inspection of the figure, it appears that the simplest way to represent the relationship is by means of a *straight line*. This is fortunate, because the method of simple correlation that is described in this chapter deals only with straight-line, or *linear*, relationships. Relationships that take the form of curved lines are measured by other methods. When it seems advisable to use a formal mathematical test to determine whether or not a relationship is linear, the description of such a test may be found in more advanced texts.[1]

Although, of course, no one line will fit all the points in Fig. 44, mathematics furnishes a formula for determining the line of best fit, which is usually called the *line of regression*

Fɪɢ. 45.—Geometric meaning of the equation of a straight line.

of Y on X. The general equation of a straight line is

$$Y_c = a_{yx} + b_{yx}X, \qquad (65)$$

where a is the *intercept* of the line on the Y axis, and b is the *slope* of the line with respect to the X axis, or the ratio of c to d in Fig. 45. (This follows from the argument that at any point, P, on the line, $Y = a + c$; but by definition $b = \dfrac{c}{X}$, or $c = bX$; therefore, $Y = a + bX$.)

To determine the values of the constants, a and b, that will give the line of best fit, the following *normal equations* are used:

[1] G. U. Yᴜʟᴇ, and M. G. Kᴇɴᴅᴀʟʟ, *An Introduction to the Theory of Statistics*, pp. 455–456, Charles Griffin & Company, Ltd., London, 1937.

$$b_{yx} = \frac{\Sigma XY - NM_xM_y}{\Sigma X^2 - NM_x^2} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma xy}{\Sigma x^2}, \quad (66)*$$

$$a_{yx} = M_y - b_{yx}M_x, \quad (67)$$

where the subscripts yx indicate the *regression of Y on X*.

From Table 50, we substitute in formula (66):

$$b_{yx} = \frac{3839.32 - 20(18.31)(10.38)}{6843.82 - 20(18.31)^2},$$

$$b_{yx} = .27516,\dagger$$

$$a_{yx} = 10.38 - .27516(18.31) = 5.34182.$$

Substituting these values of a and b in formula (65),

$$Y_c = 5.3418 + .27516X. \quad (68)$$

Putting $X = 12.5$ in formula (68), we have

$$Y_c = 5.3418 + .27516(12.5).$$
$$Y_c = 8.78130.$$

Letting $X = 22$,

$$Y_c = 11.39534.$$

Plotting these two calculated points, (12.5, 8.78) and (22, 11.395), in Fig. 44, we get the line of regression of Y on X there shown. If $X = 0$, $Y_c = 5.34 = a$.

If the origin is shifted to the means of the two series,‡ (Fig. 46), equation (65) becomes

$$y_c = bx, \quad (69)$$

where x and y are *deviates from their respective means*. For the

---

* Also, see formula (88).

† These figures are carried to several decimal places to provide a check in the summation of the third column of Table 51. If the work has been correctly done, this column will sum approximately to zero.

‡ Notice that the mean of the Y_c values calculated from the regression equation is equal to the mean of the observed Y values. This may be shown algebraically by replacing a in equation (65), above, with its equivalent from equation (67):

$$Y_c = a_{yx} + b_{yx}X,$$
$$Y_c = M_y - b_{yx}M_x + b_{yx}X,$$
$$\frac{\Sigma Y_c}{N} = M_y - b_{yx}M_x + b_{yx}M_x.$$
$$\frac{\Sigma Y_c}{N} = M_y.$$

If the second equation above is expressed in terms of mean deviates, we get

present problem, this gives

$$y_c = .27516x \tag{70}$$

which is a simpler equation and often easier to handle than equation (68). The y_c values calculated from this equation,
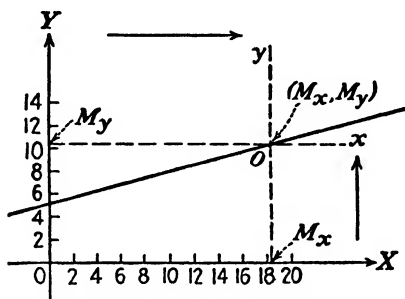
Fig. 46.—Shift of axes necessary to change regression line to mean deviate form $(y_c = bx)$.

however, are of course not directly comparable with the observed Y's. For that reason equation (68) is used to provide the values in Table 51.

A measure of the *goodness of fit* of the regression line $Y_c = 5.34 + .275X$ to the points in Fig. 44 is given by the

$$(Y_c - M_y) = (M_y - M_y) - b_{yx}(M_x - M_x) + b_{yx}(X - M_x);$$
$$(Y_c - M_y) = b_{yx}(X - M_x)$$

or

$$y_c = b_{yx}x.$$

Subtracting M_y from each Y value and M_x from each X value in equation (65) is equivalent to measuring all Y values from the mean of the Y's, and all X values from the mean of the X's. That is, the Y axis in Fig. 46 is simply moved to the right to the mean of the X's, and the X axis is moved up a distance equal to the mean of the Y's. This, of course, places the intersection of the two new axes at a point which has for its coordinates the means of the two series (M_x, M_y). Since this point is the *origin* of the system of axes from which all values of X and Y are to be measured, however, it is convenient to give it the coordinates $(0, 0)$. This is also necessary if we express x and y in mean deviate form as in equation (69), because at the point of the means the value of every mean deviate must be zero.

It follows from the second equation above that the regression line always passes through the point (M_x, M_y), since, if we let $X = M_x$,

$$Y_c = M_y - b_{yx}M_x + b_{yx}M_x, \qquad Y_c = M_y.$$

The same fact appears if we let $x = 0$ in equation (69): $y_c = b_{yx}(0)$, $y_c = 0$.

formula for the *standard error of estimate, S_y*:

$$S_y = \sqrt{\frac{\Sigma d^2}{N}}, \tag{71}$$

or

$$S_y = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}, \tag{72}$$

where d is the difference between the observed and the calculated Y values, and N is the number of paired values.   The d's are shown in Table 51:

TABLE 51.—VALUES OF d AND d^2

Observed (Y)	Calculated (Y_c)	d	d^2
9.7	10.45980	− .75980	.57730
12.0	11.45037	+ .54963	.30209
10.4	10.40476	− .00476	.00002
8.3	8.78132	− .48132	.23167
11.6	11.42286	+ .17714	.03138
6.9	10.15712	−3.25712	10.60883
10.3	10.07457	+ .22543	.05081
6.8	9.66183	−2.86183	8.19007
12.1	10.98260	+1.11740	1.24858
7.4	10.10209	−2.70209	7.30129
13.9	10.12960	+3.77040	14.21592
10.1	11.53292	−1.43292	2.05326
13.8	10.04706	+3.75294	14.08456
9.2	9.30412	− .10412	.01084
9.8	11.06515	−1.26515	1.60060
12.2	9.79941	+2.40059	5.76283
9.3	10.48731	−1.18731	1.40971
12.2	11.39534	+ .80466	.64748
10.5	10.23967	+ .26033	.06777
11.1	10.10209	+ .99791	.99582
207.6	207.59099	.00001	69.39083

$$S_y = \sqrt{\frac{69.39}{20}} = 1.86,$$

or, using formula (72),

$$S_y = \sqrt{\frac{2234.78 - 5.34182(207.6) - .27516(3839.32)}{20}} = 1.86.$$

The standard error of estimate is like the standard deviation, except that in the case of the latter the Y values are subtracted from their mean, while in the case of the former they are subtracted from the regression line, *i.e.*, from the calculated Y_c's. Notice in Table 51 that the deviations from regression add to zero, just as do mean deviations. If the distribution of Y values is normal, two out of three of the observed Y's will not vary from the regression line by more than one standard error of estimate on each side. This may be shown graphically by plotting in the range $\mp S_y$ from the regression line in Fig. 44. Adding and subtracting 1.86 and $Y_c = 8.78$ at $X = 12.5$, and then 1.86 and $Y_c = 11.40$ at $X = 22$, gives a range of 6.92–10.64 at the small end of the scale and a range of 9.54–13.26 at the large end. Accordingly, only six counties— Buffalo, Calumet, Clark, Columbia, Dane, and Douglas—out of the 20 are found to fall outside the range $\mp 1S_y$. Thus 30 per cent of the cases exceed the range, compared with 32 per cent in a strictly normal distribution. This close agreement is in spite of the small number of counties in Table 50.

There is, of course, seldom any reason for using a regression equation to calculate values of Y for comparison with the data from which the regression equation was obtained. A regression equation is rather applied to new data for the purpose of making *predictions*. For example, the usefulness of the regression equation (68), based on Table 50, lies in telling us what death rates to expect in counties that are not included in the table, or in a year other than 1935.

Even in the prediction of individual Y values when r is low, however, it is often possible to reach relatively safe conclusions by noting the odds in their favor. For example, the most probable value of Y corresponding to an X value of 18.6 was found by substituting $X = 18.6$ in equation (68), giving $Y_c = 10.46$. In other words, if we know that a county had a birth rate of 18.6, we can predict that its most probable death rate is 10.46, and we can feel some confidence that its actual death rate will not usually be below 8.60 or above 12.32 (*i.e.*, 10.46 ± 1.86). If we wish to be surer, the odds are about 20 to 1 in a normal distribution that the death rate of this county will fall between 10.46 ∓ (1.86 × 2), *i.e.*, between 6.74 and 14.18. If practical certainty is required, only once in some

369 times in a normal distribution will the death rate exceed the range of 10.46 $\mp$ (1.86 $\times$ 3), or 4.88 to 16.04, inclusive. The spread of possible error is now large, but the advantage over random guessing is still considerable.   This is usually true even after making allowance for the fact that the distribution is not normal, and for errors due to sampling.

The same principle applies to a variety of related questions, *e.g.*, What is the probability that a county with a birth rate of 17 will have a death rate as low as 8 or as high as 12? Substituting $X = 17$ in regression equation (68), we find $Y_c = 10$, approximately.   The difference between the expected death rate of 10 and a death rate of 8 or 12 is ± 2.   If we regard the death rates of all counties whose birth rate is 17 as normally distributed about a mean of 10, with a standard deviation of $Sy = 1.86$, then the difference ± 2 lies 2.00/1.86 = 1.08 standard deviation units above or below the mean.   Referring to a table of normal areas (Appendix Table 1), we see that practically 36 per cent of the area of the curve falls between the mean and an ordinate at 1.08σ.   Hence we may say that a deviation as great as or greater than 1.08σ may occur above or below the mean $100.0 - 2(36) = 28$ times in 100.   The odds are therefore 72 to 28, or roughly $2\frac{1}{2}$ to 1, against such an event.

In equations (65) and (66), *b*, which is the slope of the regression line of Y on X, is called the *regression coefficient*.   It is a useful measure, since it shows the number of Y units that the most probable value of Y changes for each unit change in X. For example, in equation (68), $Y_c = 5.34 + 0.275X$, the regression coefficient is 0.275, which means that the most probable value of Y increases 0.275 of a Y unit for every X unit that X increases.   If the equation were $Y_c = 5.34 - 0.275X$, the most probable value of Y would decrease 0.275 of a unit for each unit that X increased.

3. The Coefficient of Correlation : Ungrouped Data.—Although the table of X and Y paired values (Table 50), the scatter diagram (Fig. 44), the regression equation of Y on X (formula (65)), the regression coefficient *b*, and the standard error of estimate S_y give a great deal of information about the amount and nature of the relationship between two variables, X and Y, none of them furnishes in a single figure an index of the *amount of the relationship*.   This is supplied by the simple Pearsonian

coefficient of correlation, *r*, which for ungrouped data may be found from the following formula:

$$r = \frac{\Sigma XY - NM_x M_y}{\sqrt{(\Sigma X^2 - NM_x^2)(\Sigma Y^2 - NM_y^2)}} \tag{73}[1]$$

Applying this formula to Table 50,

$$r = \frac{3839.32 - 20(18.31)(10.38)}{\sqrt{[6843.82 - 20(18.31)^2][2234.78 - 20(10.38)^2]}},$$

$$r = \frac{38.16}{\sqrt{(138.7)(79.89)}},$$

$$r = .36.$$

Since *r* is a coefficient that can vary only from 0 to ± 1, this is not a high value, indicating rather low relationship between the birth rates and death rates in the 20 sample counties of

[1] Alternative formulas, which are sometimes convenient, are

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}, \tag{74}$$

$$r^2 = \frac{a\Sigma Y + b\Sigma XY - N\left(\frac{\Sigma Y}{N}\right)^2}{\Sigma Y^2 - N\left(\frac{\Sigma Y}{N}\right)^2}, \tag{75}$$

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}, \tag{76}$$

$$r = \frac{\Sigma XY - NM_x M_y}{N\sigma_x \sigma_y}, \tag{77}$$

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}, \tag{78}$$

$$r = \sqrt{b_{yx} b_{xy}}, \tag{79}$$

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_D^2}{2\sqrt{\sigma_x^2 \sigma_y^2}}, \tag{80}$$

where *D* refers to the differences between the raw paired values. This is known as the *difference* formula.

$$r = \frac{\Sigma xy}{N\sigma_x \sigma_y} = \frac{1}{N}\Sigma \frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y} = \frac{\Sigma x'y'}{N}, \tag{81}$$

$$r = \frac{\sigma_e}{\sigma_y}, \tag{82}$$

where σ_e is the standard deviation of the Y_e's calculated from the regression equation. See also formula (89).

Wisconsin.   It is about what would be expected from the scatter diagram (Fig. 44).

The labor of computing a correlation coefficient from ungrouped data can sometimes be reduced by dividing one or both series by some appropriate divisor, or by subtracting an arbitrary constant from the values of either or both series.   As will be seen, this does not affect the value of *r*.   The method also applies to the regression equation, provided the original values are restored.

4. Size of Sample from Which *r* Is Calculated.—It is assumed throughout the discussion of this chapter that the coefficient of correlation, *r*, is not calculated from very small numbers of paired values, say less than 25.   If this assumption is not met, and the data are regarded as a sample, many of the formulas given need correction.   Since small-sampling theory is omitted from this text, the student may see certain references listed at the end of this chapter for its treatment.[1]

5. The Meaning of the Correlation Coefficient, *r*.—It has already been seen that the standard error of estimate, S_y, around the regression line for Table 50 is approximately 1.86.   The variance of the observed *Y*'s is

$$\sigma_y{}^2 = \frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2, \tag{83}$$
$$\sigma_y{}^2 = \frac{2234.78}{20} - \left(\frac{207.6}{20}\right)^2,$$
$$\sigma_y{}^2 = 4.$$

If we compare $S_y{}^2$ with $\sigma_y{}^2$, we shall have a measure known as the *coefficient of alienation squared*, k^2:

$$k^2 = \frac{S_y{}^2}{\sigma_y{}^2} = \frac{(1.86)^2}{(2)^2} = 0.865. \tag{84}[2]$$

This shows that 86.5 per cent of the variance in county death rates remains in the form of "scatter" around the regression

---

[1] See, for example, Yule and Kendall, Ezekiel, Fisher, and Croxton and Cowden.   The student should not be misled by the circumstance that, in the example of Table 50, 20 pairs of values were treated as a large sample.   This was done only for convenience of illustration.   Strictly, small-sample methods should be used with 20 cases, although even for that size of sample it often makes no important difference.

[2] Compare the distances of the dots from the regression line and from the mean of the *Y*'s in Fig. 44.

line, which is not controlled by the birth rates. Again, by formula (78),

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2},$$
$$r^2 = 1 - k^2,$$

or

$$k^2 = 1 - r^2,$$

and

$$r^2 + k^2 = 1. \tag{85}$$

That is, r^2 and k^2 together account for 100 per cent of the variance in Y. Since we have just seen that k^2 indicates the percentage not controlled by X, $r^2 = 1 - k^2$ evidently indicates the percentage controlled by X through the medium of the regression equation. Thus, above, $r^2 = (.36)^2 = .13$, meaning that a correlation of $r = .36$ accounts for only 13 per cent of the variance of the Y series. This interpretation of r^2 is further clarified by formula (82) squared,

$$r^2 = \frac{\sigma_c^2}{\sigma_y^2}.$$

Here the numerator, σ_c^2, is the variance of the Y_c series calculated from the regression equation, so that its value is entirely controlled by X.

Substituting the values of r^2 and k^2 found in the illustrative problem above in formula (85), we get

$$(.36)^2 + .865 = .995,$$

or 99.5 per cent, the slight variation from 100 per cent being due to approximations in the calculation of r^2 and k^2.

Notice, in general, that an r as large as .71 is required to cut the variance of Y by 50 per cent (if $r^2 = .50$, then

$$r = \sqrt{.50} = .71).$$

Where both X and Y are assumed to be built up of simple elements of equal variability all of which are present in Y but some of which are lacking in X, it can be proved mathematically that r^2 measures that proportion of all the elements in Y which are also present in X. For that reason, in cases where the dependent variable is known to be causally related to the independent variable, r^2 may be called the coefficient of determination.[1]

[1] Mordecai Ezekiel, *Methods of Correlation Analysis*, p. 120, John Wiley & Sons, Inc., New York, 1930.

Although these assumptions seldom hold in practice, it is customary to regard r^2 as a better measure of relationship than r. At any rate, r^2 is a more conservative estimate.

Does the correlation between the birth rates and death rates in Table 50 mean that the birth rate is the cause of the death rate? Obviously, being born is not the cause of dying. Sanitary conditions, medical service, and various other factors determine death rates. It happens, however, that infants are more susceptible to death by disease than are older children and adults, so for this reason, other things being equal, the population with the largest proportion of infants will have the highest death rate. In general, it may be said that the presence of simple correlation between two factors may or may not be accompanied by a direct or efficient causal connection between them. Often simple correlation is due to common causes, as when teachers' salaries and the amount of money spent for alcoholic beverages rise and fall together with changes in business conditions. There is much danger that this kind of correlation will be misinterpreted. Sometimes, as in the case of the birth and death rates above, one factor is a necessary antecedent but not a direct cause of a correlated factor. Very rarely, two factors show a high but purely accidental correlation, as the yield of potatoes in Great Britain with, say, smallpox epidemics in the United States. The safest interpretation is that the presence of correlation between two factors indicates that as one increases the other tends to increase or decrease, *i.e.*, they *vary together* to some extent. *Why* they vary together may be determined by further statistical and experimental methods, such as those of partial correlation and the laboratory, which seek to control the various interfering factors involved.

Caution should be used in comparing two or more values of r. It often happens that interfering factors, of which the investigator takes no account, cause two r's that should be the same to differ widely, or two r's that should differ widely to appear the same. Unless "other things are equal," at least broadly, such comparisons have little point.

6. A Convenient Formula for the Regression Equation When r Is Known.—When the value of r is found before the regression equation is set up, the latter may conveniently be obtained from the equation

$$Y_c - M_y = r \frac{\sigma_y}{\sigma_x} (X - M_x), \tag{86}$$

or

$$y_c \doteq r \frac{\sigma_y}{\sigma_x} x. \tag{87}$$

Comparing formula (87) with formula (69), it is seen that

$$b_{yx} = r \frac{\sigma_y}{\sigma_x},$$

or

$$r = \frac{b_{yx}\sigma_x}{\sigma_y}. \tag{88}$$

7. Simple Linear Correlation Applied to Grouped Data.—The method of dealing with simple linear correlation developed above applies to ungrouped data, such as shown in Table 50. In the case of grouped data, the principles and procedures are the same, except that formulas (89) through (92) are specially adapted for use with frequency tables.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}, \tag{89}$$

where

$$\Sigma xy = \Sigma f_{xy} d_x d_y - \frac{\Sigma f_x d_x}{N} \cdot \Sigma f_y d_y, \tag{90}$$

$$\Sigma x^2 = \Sigma f_x d_x{}^2 - \frac{(\Sigma f_x d_x)^2}{N}, \tag{91}$$

$$\Sigma y^2 = \Sigma f_y d_y{}^2 - \frac{(\Sigma f_y d_y)^2}{N}, \tag{92}$$

x and y are mean deviates, d_x represents unit step deviations from an assumed mean of the X's, d_y represents unit step deviations from an assumed mean of the Y's, N is the total frequency of pairs in the table, f_x is the total frequency of pairs in an X class or column, f_y is the frequency of pairs in a Y class or row, and f_{xy} is the frequency of pairs in a cell. These symbols appear in the margins of correlation Table 53.

It is reasonable that the proportion of children in a state's population should influence the percentage of the state's income that is spent for schooling. Let us measure the extent to which this is true. The data needed are in Table 52. For our purpose it is not necessary to weight the percentage figures by the state populations.

TABLE 52.—PERCENTAGE OF POPULATION UNDER 19 YEARS OF AGE IN 1930, AND PERCENTAGE THAT SCHOOL EXPENDITURES WERE OF ALL INCOME IN 1928, BY STATES*

State	Per cent of population under 19 years of age, 1930 (X)	Per cent school expenditures were of all income, 1928 (Y)
Southeast:		
Virginia	44.4	2.61
N. Carolina	49.3	4.38
S. Carolina	50.6	3.16
Georgia	46.3	1.75
Florida	39.2	5.76
Kentucky	43.9	2.29
Tennessee	43.8	2.57
Alabama	47.0	2.74
Mississippi	46.6	3.94
Arkansas	45.3	2.55
Louisiana	44.0	2.61
Southwest:		
Oklahoma	44.2	3.27
Texas	42.6	2.57
N. Mexico	46.8	3.40
Arizona	42.1	3.67
Northeast:		
Maine	37.3	1.93
N. Hampshire	35.2	2.14
Vermont	37.0	2.24
Massachusetts	35.1	1.85
R. Island	37.0	1.89
Connecticut	37.0	2.46
N. York	33.6	2.11
N. Jersey	36.1	3.20
Delaware	35.9	1.91
Pennsylvania	39.4	2.20
Maryland	37.2	1.97
W. Virginia	46.1	3.21
Middle States:		
Ohio	36.1	3.05
Indiana	36.5	3.93
Illinois	34.9	2.28
Michigan	37.7	3.92
Wisconsin	38.0	2.95
Minnesota	38.3	3.55
Iowa	37.2	3.82
Missouri	35.7	2.46
Northwest:		
N. Dakota	45.4	6.13
S. Dakota	42.5	5.78
Nebraska	39.3	3.95
Kansas	38.1	4.24
Montana	39.0	3.96
Idaho	42.8	4.02
Wyoming	39.2	3.30
Colorado	38.0	3.29
Utah	46.1	3.91
Far West:		
Nevada	31.8	3.33
Washington	33.7	2.80
Oregon	33.1	3.31
California	30.4	3.25
United States	38.8	2.74

* From T. J. WOOFTER, JR., Landlord and Tenant on the Cotton Plantation, *WPA Research Monograph* V, 1936, p. 141.

The 48 pairs of values in Table 52 are hardly enough to justify grouping, but are convenient for illustrating the grouped method. The entries in Table 53 are made from the ungrouped data of Table 52, as follows. X represents the percentage of the population under 19 years of age, and Y is the percentage that expendi-

tures for school purposes were of total income in 1928. The first state in Table 52 has $X = 44.4$, so it will fall somewhere in col. 44.0–45.9 of Table 53. Since the corresponding Y value is 2.61, a tally is entered in row 2.40–2.79 of col. 44.0–45.9. Similarly, the second state has an X value of 49.3 and a Y value of 4.38, so a tally is placed in col. 48.0–49.9 and row 4.00–4.39 of Table 53; and so on. After all the entries are tallied in the cells, the tallies are counted and replaced by numbers.

In Table 53 we then see two ordinary frequency distributions, X and Y, placed at right angles to each other and exhibiting a double classification. The large figures in the cells are the frequencies. Instead of making a scatter diagram, as we did with ungrouped data, let us estimate the mean of the Y's in each column of the table. Consider, for example, the column with the heading 34.0–35.9. We have for the mean

$$\frac{(2.6 \times 1 + 2.2 \times 2 + 1.8 \times 2)}{5} = 2.1.$$

This may be marked by a small circle at the left side of the column, although if it did not interfere with reading the table it should be located at the mid-point of the column. Similar circles indicate the positions of the means of the other columns which have a frequency as large as five. An inspection of these means shows that they have an irregular tendency to rise in the positive direction across the table. This suggests some positive correlation between X and Y. However, the circles form more of a curve than a straight line, rising to a peak in the 38.0–39.9 column and then descending slightly. If we suppose that we are dealing with a sample thrown up by a particular set of causes, some of the irregularities may be due to random factors and a small sample. But even if we make allowance for the extreme cases in cols. 38.0–39.9, 42.0–43.9, and 44.0–45.9, the curved effect is not lessened. To assume that the relationship is linear and estimate the amount of correlation on that basis will reduce the value of the coefficient slightly, compared with the use of a coefficient of curvilinear correlation. Since we cannot deal with curvilinear correlation here, we shall use the simpler straight-line hypothesis. There is also some justification for this in view of the fact that the large scatter indicates a low correlation in any case.

Table 53.—Correlation of Percentage of Population under 19 Years of Age in 1930 (= X) with Percentage That School Expenditures Were of All Income in 1928 (= Y) in Each State of the United States*

Y \ X	30.0–31.9	32.0–33.9	34.0–35.9	36.0–37.9	38.0–39.9	40.0–41.9	42.0–43.9	44.0–45.9	46.0–47.9	48.0–49.9	50.0–51.9	(1) f_y	(2) d_y	(3) $f_y d_y$	(4) $f_y d_y^2$	(5) $\Sigma f_{xy} d_x$	(6) $d_y \Sigma f_{xy} d_x$
6.00–6.39								1				1	+8	8	64	3	24
5.60–5.99					1		1					2	+7	14	98	2	14
5.20–5.59												0	+6	0	0	0	0
4.80–5.19												0	+5	0	0	0	0
4.40–4.79												0	+4	0	0	0	0
4.00–4.39					1		1		1			3	+3	9	27	7	21
3.60–3.99		1		3	2		1		2			8	+2	16	32	7	14
3.20–3.59	2	1	2	1	3			1				10	+1	10	10	−1	−1
2.80–3.19		1		1	1							4	0	0	0	2	0
2.40–2.79			1	1			2	3	1			8	−1	−8	8	14	−14
2.00–2.39		1	2	1			2					6	−2	−12	24	−6	12
1.60–1.99			2	3	1							6	−3	−18	54	−3	9
(1) f_x	2	3	5	10	9	0	6	5	6	1	1	48		19	317	25	79
(2) d_x	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	+6						
(3) $f_x d_x$	−8	−9	−10	−10	0	0	12	15	24	5	6	25					
(4) $f_x d_x^2$	32	27	20	10	0	0	24	45	96	25	36	315					
(5) $\Sigma f_{xy} d_y$	2	−1	−11	−5	15	0	8	6	2	3	0	19					
(6) $d_x \Sigma f_{xy} d_y$	−8	3	22	5	0	0	16	18	8	15	0	79					

$$M_x = 39 + \frac{2(25)}{48} = 40$$

$$M_y = 3 + \frac{.4(19)}{48} = 3.16$$

* From T. J. Woofter, Jr., *Landlord and Tenant on the Cotton Plantation*, *WPA Research Monograph V*.

The line of regression of Y on X, and dotted lines representing $\pm 1S_y$, the values for which are worked out below, are drawn in the correlation table (Table 53). A study of them in relation to the entries in the correlation table should be helpful, just as it was in the case of the scatter diagram for ungrouped data (see Fig. 44). It appears from Table 53 that the actual relationship changes from strongly positive in the left half of the table to moderately negative in the right half, whereas the linear regression implies a constant positive correlation throughout. Also, the linear equation is far from fitting the data of the two halves of the table equally well. On the other hand, in only one column does the proportion of items falling outside the range of one standard error of estimate around the regression line exceed the normal one-third. In practice it would probably not be worth while to carry the analysis any farther. We shall, however, use the table to show the steps involved in calculating the Pearsonian correlation coefficient, r, the linear regression of Y on X,[1] the standard error of estimate, S_y, and other statistics, from grouped data.

Proceeding with Table 53, we enter unit-step deviations in row (2) and col. (2). The entries in row (3) and col. (3) and in row (4) and col. (4) are familiar and should be obvious from the symbols. Next, we multiply each cell frequency first by d_x and place the product in the upper right-hand corner of the cell, and then by d_y and place the product in the lower left-hand corner of the cell. The d_x products are then added by rows and the d_y products by columns. Column (6) and row (6) are obtained by multiplying the entries in col. (5) and row (5) by d_y and d_x, respectively, and the products are summed over the column and the row.[2]

We finally substitute from Table 53 in formulas (90)–(92),

$$\Sigma xy = 79 - (19)\tfrac{25}{48} = 69.1,$$

$$\Sigma x^2 = 315 - \frac{(25)^2}{48} = 302,$$

$$\Sigma y^2 = 317 - \frac{(19)^2}{48} = 309.5,$$

[1] There is also always a regression line of X on Y, from which the most probable values of X may be calculated for given values of Y. The two regression lines are not the same. To find the regression of X on Y, simply change places with X and Y in the equations given in this chapter.

[2] As a check on the work, notice that in Table 53 cols. (3), (5), and (6) should have the same totals as rows (5), (3), and (6), respectively.

Substituting these values in formula (89),

$$r = \frac{69.1}{\sqrt{(302)(309.5)}} = \frac{69.1}{\sqrt{93469}} = \frac{69.1}{305.7},$$

$$r = .23.$$

This value of r indicates very little relationship. Nevertheless, for purposes of demonstration, we shall show the use of the formulas for finding the regression equation of Y on X and the coefficient of alienation, k. We have

$$b_{yx}' = \frac{\Sigma xy}{\Sigma x^2} \tag{93}$$

$$b_{yx}' = \frac{69.1}{302} = 0.23.$$

But this value of b is in terms of unit-step deviations or class intervals. To change it back to scale units,

$$b_{yx} = \frac{i_y b_{yx}'}{i_x}, \tag{94}$$

where i_y = class interval of Y.
     i_x = class interval of X.

$$b_{yx} = .23 \frac{(.4)}{2} = 0.046.$$
$$a_{yx} = M_y - b_{yx}M_x. \tag{95}$$
$$a_{yx} = 3.16 - .046(40).$$
$$a_{yx} = 1.32.$$

Therefore, substituting in formula (65), we have

$$Y_c = 1.32 + .046X$$

also,

$$S_y^2 = \sigma_y^2(1 - r^2), \tag{96}$$

$$\sigma_y^2 = i_y^2 \left[\frac{\Sigma f_y d_y^2}{N} - \left(\frac{\Sigma f_y d_y}{N} \right)^2 \right], \tag{97}$$

$$\sigma_y^2 = (.4)^2 \left[\tfrac{317}{48} - (\tfrac{19}{48})^2 \right],$$
$$\sigma_y^2 = 1.03,$$
$$S_y^2 = 1.03(1 - .0529),$$
$$S_y^2 = .9755,$$
$$k^2 = \frac{S_y^2}{\sigma_y^2} = \frac{.9755}{1.03} = .95.$$

Thus an r of .23 leaves 95 per cent of σ_y^2 as scatter around the

regression equation, or improves prediction only

$$r^2 = 1 - k^2 = .05,$$

or about 5 per cent, in terms of the variance of the Y's.

The student is asked to check the plotting of the regression line and the lines showing the standard error of estimate in Table 53.

Regression equations (69), (86), and (87) also apply to grouped data.

From the above, it is clear that there is little tendency for the percentage of income expended for schools to be proportionate to the percentage of children under 19 years old in the population when states are taken as units and a linear relationship is assumed. Apart from the latter assumption, which has already been discussed, it may well be objected that a state is a large area, within which very different relations between these two percentages may exist. Thus a large city and a rural county in the same state may be more sharply unlike in this respect than two cities in separate states. For this reason, the average relationship given for each state as a whole is likely to be unrepresentative, and so to lack meaning. It would be much better if the data were available by school districts, in which case a higher correlation might be found.

8. The Rank Correlation.—A method of linear correlation that takes account of the rank orders of paired items but disregards their values is sometimes used for rough work, or when the values of the items are not known. The formula is

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}, \tag{98}^1$$

where D is the difference between the ranks of a pair of items, and N is the number of pairs.

As an illustration of the use of this formula, let us refer back to Table 50, and rank the counties with respect to their death rates and birth rates, as shown in Table 54. When there are ties, as between Douglas and Eau Claire counties in death rates, and Clark and Fond du Lac in birth rates, the tied items are given the mean of the ranks they would occupy if they were not equal, and the next item takes the rank just above the highest

---

[1] ρ is the lower-case Greek letter rho.

rank used in finding the tied mean. For example, the ranks 7 and 8 are averaged to give $\frac{7+8}{2} = 7.5$ as the mean rank of Clark and Fond du Lac counties, and Columbia county has the rank 9.

TABLE 54.—TWENTY WISCONSIN COUNTIES RANKED WITH RESPECT TO BIRTH RATES AND DEATH RATES (LOW TO HIGH)

County	Rank in birth rate (X)	Rank in death rate (Y)	Difference (D)	D²
Bayfield	1	4	− 3	9
Dodge	2	5	− 3	9
Calumet	3	1	2	4
Douglas	4	17.5	−13.5	182.25
Dane	5	19	−14	196
Burnett	6	10	− 4	16
Clark	7.5	3	4.5	20.25
Fond du Lac	7.5	13	− 5.5	30.25
Columbia	9	20	−11	121
Buffalo	10	2	8	64
Florence	11	12	− 1	1
Barron	12	11	1	1
Adams	13	7	6	36
Dunn	14	6	8	64
Chippewa	15	16	− 1	1
Door	16	8	8	64
Eau Claire	17	17.5	− .5	.25
Brown	18	14	4	16
Ashland	19	15	4	16
Crawford	20	9	11	121
Total	....	....	......	972

Substituting in formula (98),

$$\rho = 1 - \frac{6(972)}{20(400 - 1)} = .27.$$

Like r, the value of ρ may vary from $+1.0$ to -1.0.

Exercises

1. a. What is the amount of relationship between the length of French and English words in the accompanying table? Plot the data, and discuss the scatter diagram. Is the relationship reasonably linear? Use both the ungrouped and the grouped methods of calculating r as a

check. Do the two methods necessarily give exactly the same value of
r? Explain. Just what does r mean in this case?

NUMBER OF LETTERS IN A SAMPLE OF FRENCH WORDS (X), AND IN THEIR
NEAREST ENGLISH EQUIVALENTS (Y)

X	Y	X	Y	X	Y	X	Y
1	2	4	5	6	7	8	8
9	9	6	6	2	4	9	10
8	8	6	4	3	6	5	5
6	7	8	9	8	7	9	9
4	7	5	10	7	7	4	5
7	7	4	4	5	5	5	4
7	7	10	17	6	6	3	3
6	6	5	6	5	4	6	5
8	11	8	9	8	10	12	11
8	8	5	6	4	6	5	5
8	11	5	4	11	8	8	9
8	7	3	3	5	3	11	11
8	7	4	3	10	10	8	8
9	8	7	5	8	8	7	6
7	5	7	8	11	8	7	8
5	5	6	7	10	8	7	7
6	5	9	7	8	10	7	7
12	9	8	6	11	9	6	5
5	8	5	4	10	8	10	9
8	7	6	2	10	7	9	9
10	10	5	4	7	6	9	8
7	8	8	8	11	10	8	9
5	8	6	7	9	9	9	11
8	9	9	9	13	6	8	7
7	8	9	8	8	9	5	6
10	11	8	7	9	7	7	7
8	8	3	6	8	10	7	8
3	3	11	17	9	10	8	5
7	7	5	4	6	7	5	6

b. Get the regression of Y on X from both the ungrouped and the
grouped data, as a check, plot the line, and explain what a and b mean
in the equation.

c. What is the most probable length of an English word correspond-
ing to a French word of six letters?

d. Within what range will the number of letters in the English words in (*a*) fall two times out of three? Ninety-five times out of 100? Practically always?

e. What is the value of the coefficient of alienation squared, and what does it mean here?

f. What is the coefficient of determination and its interpretation in this problem?

g. Find the coefficient of rank correlation, ρ, for the same data, and compare its value, meaning, and adequacy with r.

2. For the table below, find the value of r and of b, and compare them in meaning.

AGE OF FATHERS (*Y*) CORRELATED WITH AGE OF SONS (*X*)

Y \ X	25	27	29
60	3	5	7
65	2	11	14
70		2	6

3. Find r^2 and k^2 for the following table, and explain their meaning.

NUMBER OF CHILDREN IN THE FIRST GENERATION OF SIX FAMILIES (*X*), AND THE AVERAGE NUMBER OF CHILDREN IN THE SECOND GENERATION OF THE SAME FAMILIES (*Y*)

X	3	4	6	7	9	15
Y	3	2	4	4	5	5

4. *a.* By inspection, is there any relationship between the votes of the states in 1876 and in 1932? If any, is it positive or negative?

REPUBLICAN VOTE FOR PRESIDENT IN NINE STATES, 1876 AND 1932

State	Per cent of vote Republican	
	1876	1932
Massachusetts.............	58	48
New York................	48	41
Wisconsin................	55	32
Missouri.................	41	35
Virginia.................	41	30
Mississippi..............	21	4
Louisiana................	48	7
Nevada..................	53	31
California...............	51	39

b. What does the scatter diagram show?

c. What is the equation of the regression of Y on X, where X is the percentage of the vote Republican in 1876, and Y is the percentage of the vote Republican in 1932?   Plot the line in the scatter diagram.

d. What is the standard error of estimate?   Plot it in the scatter diagram.

e. What is the most probable percentage of the vote Republican in 1932 of a state that voted 55 per cent Republican in 1876?

f. Assuming a normal distribution about the regression line, within what limits of error will the percentage vote fall two out of three times?   20 out of 21 times?   Within what limits of error does it actually fall in each case?

5. *a.* What does the scatter diagram show in the case of the accompanying table of death rates in Connecticut and Massachusetts?

<div align="center">DEATH RATE IN CONNECTICUT AND MASSACHUSETTS*</div>

Year	1924	1923	1922	1921	1920	1919	1918
Connecticut	11.3	12.0	12.0	11.4	13.6	13.3	20.4
Massachusetts	12.0	13.0	12.8	12.2	13.8	13.6	20.9

* From B. H. CAMP, *The Mathematical Part of Elementary Statistics*, p. 144.   D. C. Heath & Company, Boston, 1935.

b. If the death rate for Massachusetts is 12 in 1924, what is the most probable death rate for Connecticut in the same year in terms of the relationship between the two?

c. How much of the variance still remains as scatter in predicting a death rate in Connecticut from one in Massachusetts?

<div align="center">**References**</div>

CHADDOCK, R. E.: *Principles and Methods of Statistics*, Chap. XII, Houghton Mifflin Company, Boston, 1925.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chap. XXII, Prentice-Hall, Inc., New York, 1939.

DAVIES, G. R., and DALE YODER: *Business Statistics*, Chap. VI, John Wiley & Sons, Inc., New York, 1937.

EZEKIEL, MORDECAI: *Methods of Correlation Analysis*, Chaps. III–V, VII–IX, John Wiley & Sons, Inc., New York, 1930.

FISHER, R. A.: *Statistical Methods for Research Workers*, 4th ed., Chap. VI, Oliver and Boyd, Edinburgh, 1932.

GARRETT, H. E.: *Statistics in Psychology and Education*, Chap. IV, Longmans, Green & Company, New York, 1926.

LINDQUIST, E. F.: *A First Course in Statistics*, Chap. XI, Houghton Mifflin Company, Boston. 1938.

MILLS, F. C.: *Statistical Methods,* rev. ed., Chap. X, Henry Holt and Company, Inc., New York, 1938.

THURSTONE, L. L.: *Fundamentals of Statistics,* Chaps. XXII–XXIV, The Macmillan Company, New York, 1925.

TRELOAR, ALAN E.: *Elements of Statistical Reasoning,* Chaps. VII and VIII, John Wiley and Sons, New York, 1939.

WHITE, R. C.: *Social Statistics,* Chap. XI, Harper & Brothers, New York, 1933.

YULE, G. K., and M. G. KENDALL: *An Introduction to the Theory of Statistics,* Chaps. XI–XIII, XV, XVI, Charles Griffin & Company, Ltd., London, 1937.

CHAPTER XI

GROSS RELATIONSHIP BETWEEN TWO FACTORS: NONQUANTITATIVE CORRELATION

1. Qualitative Data.—The method of correlation described up to this point has dealt with quantitative series only, *e.g.*, birth and death rates, and proportion of state income spent for education. It often happens in sociological investigations, however, that it is needed to know the amount of relationship between two factors, one or both of which are qualitative. Examples of qualitative factors are rural or urban residence; personality ratings like Annoying, Unsympathetic, Sympathetic; occupational classes—Professional, Proprietor, Clerical, Skilled, Unskilled; and so on. Methods for correlating data of this type have been devised. Before using them, effort should be made to convert the qualitative attributes into quantitative variables, because the latter are usually more accurate and reliable. Thus, a student might be classified by the number of credits earned in college, rather than as Sophomore or Junior.

2. Reliability of Classification.—Since much depends on the reliability with which the nonquantitative variables are classified, it is advisable to have the classification repeated by two or more qualified persons. If the results are very different, better criteria for classification should be developed, or the problem dropped.

This point may be illustrated. The questionnaire that the members of a class in statistics filled out regarding their previous training in mathematics called for the sex of each student. If it were desired to correlate success in mathematics with sex, the members of the class might be divided by sex, and then subdivided into, say, four groups according to the average grades received in mathematics. This would give a table like Table 55.

The question of the reliability of the classification by class standing in this table can be dismissed, because it is based on a

quantitative variable, the average grades received in mathematics. The sex classification might be somewhat unreliable if it depended merely on the Christian names of the students in the questionnaires; but reference to the questionnaire used shows that the students checked the words Male and Female. The reliability of this classification can therefore also be accepted with confidence. We may then proceed to find the amount of relationship between the two factors in the table.

TABLE 55.—STUDENTS IN A STATISTICS CLASS GROUPED BY SEX AND GRADES RECEIVED IN MATHEMATICS

Sex	Students by class standing				
	1	2	3	4	Total
Male...........................	4	6	6	3	19
Female.........................	7	13	15	11	46
Total........................	11	19	21	14	65

All classifications are not so simple as those in Table 55, however. In Table 56, for example, a second competent person classified only 66 cases out of each 100 in the same way that this table shows, with respect to the economic status of the family. This was considered sufficient reason for abandoning the table.

TABLE 56.—ECONOMIC STATUS OF THE FAMILY IN WHICH PAROLEE WAS REARED AND OUTCOME ON PAROLE

Status	Parolees	Parole violators	
		Number	Per cent
Poor.............................	287	44	15.3
Moderate.........................	261	26	10.0
Comfortable......................	59	6	10.2
Unknown..........................	22	3	——*
Total........................	629	79	

* Sample too small to warrant an estimate.

3. Choice of a Method.—After the reliability of the classifications in a nonquantitative correlation table has been established,

the question of how to calculate the amount of relationship between the two factors in the table arises. The answer depends on the nature of the particular factors to be correlated. It is convenient to set up a key, as in Table 57, which will suggest what method should be used in each case.

The terms in Table 57 need definition and illustration. *Quantitative* means expressed in countable units, as crime rates or heights of male freshmen. *Qualitative* refers to nonmeasured traits, like those mentioned in the first paragraph of this chapter. *Qualitative Ordered* refers to qualitative categories that can be arranged in ascending or descending order, as Favorable, Indifferent, Hostile. *Qualitative Unordered* applies to qualitative categories that cannot be arranged in ascending or descending order, *e.g.*, Law, Medicine, Engineering. A *dichotomous* series is a series of two mutually exclusive and exhaustive categories, as Good, Not Good; Sick, Not Sick; Male, Female; College Graduates, Others; Families with Less than Four Children, Families with Four or More Children.

TABLE 57.—KEY TO SELECTED METHODS OF NONQUANTITATIVE CORRELATION

Variable *A*	Variable *B*	Method
Quantitative: several classes	Dichotomous	Biserial, r_{bis}
Quantitative or qualitative: ordered or unordered; several classes	Qualitative: ordered or unordered; several classes or dichotomous	Contingency, C
Dichotomous	Dichotomous	Tetrachoric, r_t Yule's Q Fourfold r_4

It is not feasible to deal here with more than the five methods listed in Table 57, though a number of less prominent methods are omitted.

4. Biserial Correlation.—In a study of divorce data for the United States in 1929, it is desirable to know whether there is any correlation between the party to whom the divorce was granted and the number of children affected. The data are shown in the first four columns of Table 58. We have here a

quantitative series to be correlated with a dichotomous series. According to the key in Table 57, this requires the biserial method of correlation.

The biserial method that we shall employ assumes that the dichotomous trait is normally distributed and continuous (*i.e.*, there is no gap in the series, and no disarrangement of an ordered series). The relationship must be linear, or that of a straight line. In the present case the idea of normality at first seems to have little meaning. However, if we think of the possibility of measuring the extent to which the husband or the wife is responsible for the granting of the divorce, and if it is reasonable to suppose that one party will seldom be wholly the instigator, but that in most cases both will be about equally involved, we may perhaps assume that the distribution of the dichotomous factor is fairly normal.
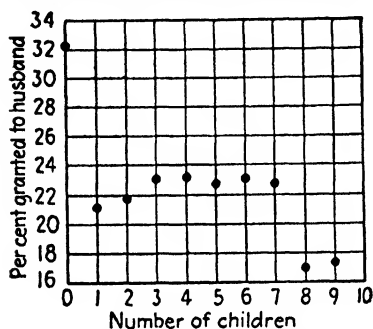
Fig. 47.—Relation of percentage of divorces granted to husband and number of children involved.

Since all reported divorces are included, the series is continuous. As a rough test whether or not the relationship is linear, the scatter diagram shown in Fig. 47 is used. In this figure are plotted the percentages of the divorces granted to husbands by the number of children affected. The trend, if any, is very irregular. Where there are no children, a much larger percentage of divorces is granted to the husband than where there are children. When the number of children is very large—*i.e.*, eight or nine—the proportion of divorces granted to the husband falls to a minimum. When the number of children affected ranges from one to seven, the proportion of divorces granted to the husband remains practically stationary. The low percentages of divorces granted to the husband when there are eight or nine children may be unreliable because of the small number of cases involved; but the circumstance that the percentage is low both for eight children and for nine children tends to support the observed figures. There seems to be little reason for calculating the value of r_{bi} in this case. We shall do so merely to show the method.

TABLE 58.—DIVORCES GRANTED, CLASSIFIED ACCORDING TO NUMBER OF
CHILDREN AFFECTED: 1929*

Children affected	Divorces granted			
	To husband	To wife	Total (f)	Per cent to husband (q_x)
(1)	(2)	(3)	(4)	(5)
0	36,840	76,970	113,810	.3237
1	8,385	32,223	40,608	.2065
2	4,255	15,242	19,497	.2182
3	1,841	6,161	8,002	´ .2301
4	774	2,571	3,345 ,	.2314
5	352	1,191	1,543	.2281
6	155	518	673	.2303
7	68	245	313	.2173
8	22	108	130	.1692
9	16	77	93	.1720
Total............	52,708	135,306	188,014	.2803
Mean............	0.55	0.77	0.71	

* From *Marriage and Divorce*, 1929, p. 41, U. S. Bureau of the Census. "Nine or more children" taken as nine, and "no report as to children" disregarded.

Apparently, the relationship in Table 58 is not linear. We shall work out the correlation, however, on the assumption that it is linear. The difference is unimportant here.

The formula for finding biserial r is

$$r_{\text{bis}} = \frac{\bar{M}_2 - \bar{m}_1}{\sigma} \frac{(pq)}{y}. \tag{99}$$

where $\bar{m}_1$ is the mean of the smaller frequency distribution [cols. (1) and (2) of Table 58], $\bar{M}_2$ is the mean of the larger frequency distribution [cols. (1) and (3)], σ is the standard deviation of the total frequency distribution [cols. (1) and (4)], p is the proportion that the total frequency of the larger distribution [col. (3)] is of the grand total frequency [col. (4)], $q = 1 - p$, and y is the height of the ordinate of a normal curve of unit area and unit standard deviation at the point separating the area of the curve into the proportions p and q, as found from Appendix Table 1. The means and standard deviation required are calculated by the usual method of unit-step deviations from an assumed mean. The required values are

$$\bar{m}_1 = 0.55,$$
$$\bar{M}_2 = 0.77,$$
$$\sigma = 1.14,$$
$$p = \frac{135,306}{188,014} = .72,$$
$$q = 1.00 - .72 = .28,$$

To find y, we turn to Appendix Table 1.   In Fig. 48 a normal curve is shown.   As explained elsewhere, the values given in the body of the table represent the proportion of the area of the curve included between the mean ordinate (shown at zero in the figure) and ordinates erected at various distances, measured in standard deviation units, from the mean.   Since here $p = .72$,
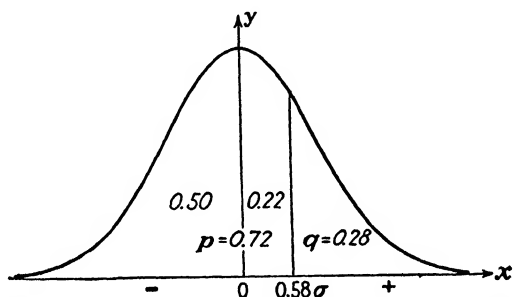
FIG. 48.—Normal curve used to find value of y in formula (99).

we need to find the height of the ordinate which divides the curve so that .72 of its area falls to the left and .28 to the right. Evidently, .72 of the area will occupy the whole left half of the curve, and a proportion $.72 - .50 = .22$ will extend into the right half.   Looking for .22 in the column of the table headed "Area," we find as the nearest approximation to it the figure 0.2190, and note that the corresponding figure in the column headed "Ordinate (y)" is 0.3372.   We therefore have $y = 0.3372$, and are ready to substitute in formula (99):

$$r_{bis} = \left(\frac{0.77 - 0.55}{1.14}\right) \frac{(.72)(.28)}{.3372}.$$
$$r_{bis} = .12.$$

As would be expected from our preliminary analysis of Table 58 and Fig. 47, the amount of linear relationship between divorces granted to husbands and the number of children affected is very slight.

The sign of r_{bis} indicates the direction of the relationship between the quantitative factor and the proportion of cases in the distribution represented by p in formula (99). Here there is a slight positive association between number of children and divorces granted to the wife, or a slight negative association between number of children and divorces granted to the husband.

The general conclusion from this analysis is that, if any correlation is present at all, there is a very slight tendency for the husband to receive the divorce relatively less often as the number of children increases. Much more informative, however, was the interpretation made from the scatter diagram in Fig. 47, that the proportion of divorces granted to the husband (1) was considerably greater where there were no children at all, (2) was little affected by increases in the number of children from one to seven, and (3) was a minimum when the number of children was eight or more.

Biserial correlation is a special adaptation of the method of correlation used in finding the Pearsonian coefficient of correlation, r, for quantitative data. For this reason, r_{bis} may be regarded as the nearest approximation to r that can be found when a quantitative series is correlated with a dichotomous series.

5. The Coefficient of Contingency.—A total of 1,118 inmates of a state prison were classified as murderers, sex offenders, and property offenders. It was wanted to know how much, if any, correlation existed between these three criminal types and intelligence. An intelligence test was given to all the men, with the results shown in Table 59. This table contains one quantitative series and one unordered qualitative classification. The key in Table 57 indicates the method of contingency for finding the amount of association present. This coefficient is based on the Chi-square (χ^2) method, and measures the amount of deviation of the observed frequencies in the table from purely random or chance frequencies. The method of finding the chance or theoretical frequencies, f_t, is based on two elementary theorems in the mathematics of probability which have already been treated (see Chap. IX). Thus, the probability that any criminal will fall in, say, the first column of Table 59 is the ratio of the total number that fall in that column to the total frequency

TABLE 59.—A PRISON POPULATION CLASSIFIED BY TYPE AND BY INTELLIGENCE QUOTIENT*

	I.Q.	Murderers					Sex offenders					Property offenders					Total
		(1) f_o	(2) f_t	(3) $f_o - f_t$	(4) $(f_o - f_t)^2$	(5) $\frac{(f_o - f_t)^2}{f_t}$	(6) f_o	(7) f_t	(8) $f_o - f_t$	(9) $(f_o - f_t)^2$	(10) $\frac{(f_o - f_t)^2}{f_t}$	(11) f_o	(12) f_t	(13) $f_o - f_t$	(14) $(f_o - f_t)^2$	(15) $\frac{(f_o - f_t)^2}{f_t}$	(18) m
(1)	40- 49	[4]	[1.0644]				[3]					[10]					[17]
(2)	50- 59	[11] 15	[5.2594] 6.3238	8.6762	75.273	11.903	[16] 19	11.1117	7.8883	62.221	5.600	[57] 67	83.5644	−16.5644	274.379	3.283	[84] 101
(3)	60- 69	13	13.7120	−.7120	.507	0.037	41	24.0937	16.9063	285.823	11.863	165	181.1945	−16.1940	262.246	1.447	219
(4)	70- 79	15	18.3453	−3.3453	11.191	0.610	25	32.2350	−7.2350	52.345	1.624	253	242.4194	10.5806	111.949	0.462	293
(5)	80- 89	14	15.5277	−1.5277	2.335	0.150	15	27.2842	−12.2842	150.902	5.531	219	205.1878	13.8122	190.777	0.930	248
(6)	90- 99	3	9.5796	−6.5796	43.291	4.519	13	16.8326	−3.8326	14.688	.873	137	126.5876	10.4124	108.418	0.856	153
(7)	100-109	[8] 10	[4.7585] 6.5116	3.4894	12.166	1.868	[5] 10	11.4418	−1.4418	2.079	.182	[63] 84	[62.8802] 86.0465	2.0465	4.188	0.049	[76] 104
(8)	110-119	[2]	[1.6279]				[5]					[19]	[21.5116]				[26]
(9)	120-129	[0]	[0.1252]				[0]					[2]	[1.6547]				[2]
(10)	Total m	70				19.087	123				25.673	925				7.230	1,118

* Adapted from an unpublished study by J. L. Gillin. f_o = observed frequency; f_t = expected frequency.

of the table, or $n_i/N = 70/1,118$, where n_i is the total frequency in col. (1), and N is the total frequency of the table. Likewise, the probability that any criminal will fall in, say, the first row is the ratio of the total number that fall in that row to the total frequency of the table, or $_1n = 17/1,118$. Now the probability of two independent events occurring together is the product of the probabilities of their separate occurrences. Therefore, the probability that any criminal will fall in both the first column and the first row of the table is

$$\left(\frac{n_1}{N}\right)\left(\frac{_1n}{N}\right) = \frac{_1nn_1}{N^2} = \left(\frac{70}{1,118}\right)\left(\frac{17}{1,118}\right) = 0.000952.$$

This means that about one out of every 1,000 prisoners in Table 59 may be expected by chance alone to fall in the cell common to col. (1) and row (1). Since there are 1,118 prisoners in the table, the expected frequency is

$$\frac{_1nn_1(N)}{N^2} = \frac{17(70)}{(1,118)^2}(1,118) = 1.0644.$$

This formula may evidently be shortened, however, to

$$f_t = \frac{_i nn_i}{N}, \qquad\qquad (100)$$

giving for the above $f_t = \dfrac{17(70)}{1,118} = 1.0644$, again. We now write this expected frequency in row (1) and col. (2) of Table 59. By use of formula (100), all of the expected frequencies are calculated and entered in cols. (2), (7), and (12). This computation is more easily done for any column by setting n_i/N in the calculating machine and multiplying it successively by the total row frequencies, $_1n$. It is a general principle of the χ^2 test that no cell should contain much less than five expected frequencies. Any cell that offends in this respect should be combined with the cell above or below it. For this reason, in Table 59 the frequencies of the first row and of the last two rows are combined with those just below or above. Comparing now the observed with the theoretical frequencies, we notice a considerable amount of difference. This indicates some association between the criminal classifications and intelligence. We proceed to measure it by computing χ^2:

$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t} \tag{101}$$

$$\chi^2 = 19.087 + 25.673 + 7.230 = 51.990$$

Substituting in the formula for the coefficient of contingency, C,

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}, \tag{102}$$

$$C = \sqrt{\frac{51.99}{1118 + 51.99}} = \sqrt{.0444},$$

$$C = .21$$

The amount of association between the types of criminals and intelligence is seen to be low. If we regard our 1,118 prisoners as a random sample, what is the probability that the value of C is zero in the total population from which it was drawn? Before we can refer this question to a table of χ^2 (Appendix Table 2), we must have regard for the proper degrees of freedom. It will be recalled[1] that in each row and column of a contingency table (*e.g.*, Table 59), one of the cell frequencies is not "free," because it may be determined by subtraction from the marginal totals. In any row or column, therefore, the number of free cell frequencies, or degrees of freedom, is one less than the number of cells (columns or rows). In Table 59 there are three columns and six rows, so that the degrees of freedom for the whole table are $(3 - 1)(6 - 1) = (2)(5) = 10$. With 10 degrees of freedom, we find in Appendix Table 2 that a χ^2 as great as 23 would occur by chance only once in 100 trials. Since our $\chi^2 = 52$ is still larger, we can be sure that the differences are not random. That is equivalent to saying that the value of C indicates a low but genuine association between types of criminals and intelligence.

C is usually found from a shorter formula than that used above:

$$C = \sqrt{\frac{S - 1}{S}}, \tag{103}[2]$$

where

$$S = \sum \left(\frac{f_0^2}{nn_i} \right), \tag{104}$$

[1] See Chap. IX, p. 148.
[2] For the derivation of this formula, see Karl J. Holzinger, *Statistical Methods for Students in Education*, p. 275, Ginn and Company, Boston, 1928.

The value in parentheses in (104) is calculated for each cell of the table, and these cell values are summed over the table. Thus for the cell 80–89 in col. (6),

$$\left(\frac{f_o{}^2}{_jnn_i}\right) = \frac{(15)^2}{248(123)} = .0074.$$

Formula (103), however, does not provide a value of χ^2 by which to test the significance of the association found.

The coefficient of contingency has the defect that it understates the amount of correlation actually present, in inverse proportion to the number of cells in the table. For a 3×3 table having perfect correlation, C would not be 1.00, as it should, but .816; for a 5×5 table, the maximum value of C is .894; for a 7×7 table, .926; for a 10×10 table, .949. Evidently C is not comparable between tables with different numbers of cells. For these reasons, it is well to apply C only to tables having say from 25 to 100 cells.

It is possible to correct C to some extent for the above fault in cases where the correlation table has a fairly normal surface, as shown by the row and column totals in ordered series. For this purpose Table 60 may be used in connection with formula (105):

$$\bar{C} = \frac{C}{t_r t_c}. \tag{105}$$

If for the moment we regard Table 59 as normal, we have from Table 60 for three columns $t_c = .859$, and for six rows $t_r = .959$, so that

$$\bar{C} = \frac{.21}{.959(.859)} = .25.$$

TABLE 60.—FACTORS FOR CORRECTING C FOR BROAD GROUPING*

Number	Correction Factor (t_r, t_c)	Number	Correction Factor (t_r, t_c)
2	.798	9	.981
3	.859	10	.985
4	.915	11	.987
5	.943	12	.989
6	.959	13	.991
7	.970	14	.992
8	.976	15	.993

The change in the value of C in this case is slight, and will always be so where the original value of C is low.   The correction is therefore worth making only when the value of C is fairly high.   Moreover, in the present case, one of the series in Table 59 is unordered, so we are not justified in regarding it as approximately normal in form, or in applying this correction to the C obtained from it.

A coefficient of contingency, C, needs perhaps even more careful interpretation than other coefficients of correlation.   In the first place, it has no sign, so that its meaning is dependent upon an examination of the correlation table itself.   When both series are ordered, it is possible to assign a sign to C; otherwise, not.   In Table 59, the prisoner classification is unordered, so the C we found can have no sign.   Notice also that the sizes of the χ^2's for the three classes of criminals are not comparable, because the number of prisoners is different in each class.   We may, however, compute the mean I.Q. for each of the three classes, and in that way note how they compare in intelligence.   Thus we find that property offenders are most intelligent with an I.Q. of 79.96, while murderers and sex offenders are approximately equal with I.Q.'s of 75.71 and 74.51, respectively.   If the categories were Life Sentence, Medium Sentence, Short Sentence, instead of Murderers, Sex Offenders, Property Offenders, the sign of C might be regarded as negative, since intelligence increases as the length of prison sentence decreases.   If neither factor in the table was quantitative, means could not be computed.   In that case, we could only compare the columns with respect to the proportions of their frequencies falling in each category of the stub.

6. Correlation in Fourfold Tables.—Any scale may be divided into just two parts, or *dichotomies*.   For example, we may measure head lengths, and then classify heads below a certain length as short, and those of this length and above as long.   Many sociological variables that have never been measured are commonly treated as dichotomies, *e.g.*, Cooperative, Not Cooperative.   Some information is gained if a more detailed breakdown is feasible, such as Completely Cooperative, Very Cooperative, Average Cooperative, Uncooperative, Completely Uncooperative.

Some qualities are most conveniently regarded as attributes rather than as quantitative variables, and naturally take a

dichotomous form. Examples are Violator of Parole, Non-violator of Parole; White Race, Other Race.

The measurement of the amount of relationship in a 2×2 table is usually rather rough and inexact, regardless of what method is used. On this account, such a table is often merely tested for the presence of relationship, without attempting to measure it. The Chi-square test, explained in Chap. IX, is commonly relied on for this purpose.

Suppose we are interested in whether or not there was any association between the occupation of agriculture and the tendency to commit crime in a given state over the period 1920–1930. Table 61 gives all the information at hand bearing on the question, together with the scheme of symbols used in a short formula for χ^2 adapted to a 2×2 table.

TABLE 61.—OCCUPATIONAL DISTRIBUTION OF THE ADULT MALE PRISON AND NONPRISON POPULATIONS OF A GIVEN STATE, 1920–1930

Occupational classification	Mean prison population	Mean nonprison population	Total
Agriculture.............	690 (u)	1,100,000 (v)	1,100,690 $(_1n)$
Nonagriculture..........	2,310 (w)	900,000 (x)	902,310 $(_2n)$
Total...............	3,000 (n_1)	2,000,000 (n_2)	2,003,000 (N)

$$\chi^2 = \frac{(ux - vw)^2 N}{n_1 n_2 (_1n_2n)}, \tag{106}$$

Substituting in this formula,

$$\chi^2 = \frac{[(690)(900,000) - (1,100,000)(2,310)]^2 2,003,000}{(3,000)(2,000,000)(1,100,690)(902,310)},$$
$$\chi^2 = 1,239.$$

Entering Appendix Table 2 with one degree of freedom, as we did in Chap. IX, we see that so large a value of χ^2 would occur by chance much less often than once in 100 times. We may, therefore, regard the presence of association between the occupation of agriculture and the commitment of crime in Table 61 as established beyond doubt.

If it seems worth while to go farther than the χ^2 test, and try to estimate approximately the degree of association in a 2×2

table, there are several coefficients available. They are based on different principles, however, and give different results. We shall illustrate three such coefficients, namely, Yule's Q, the ordinary coefficient of correlation adapted to fourfold tables, r_4, and the coefficient of tetrachoric correlation, r_t. Where one of them will not meet the needs of a particular problem, another usually will.

The formula for Yule's Q is $Q = \dfrac{ux - vw}{ux + vw}$, (107)

where the symbols refer to cell frequencies as shown in Table 61. Let us apply it to the data of Table 61. Substituting, we have

$$Q = \frac{(690)(900,000) - (1,100,000)(2,310)}{(690)(900,000) + (1,100,000)(2,310)},$$
$$Q = -.61.$$

According to this coefficient, there is a moderate amount of negative association between the occupation of agriculture and imprisonment for crime in Table 61, or, more generally, between the first column and the first row factors, when the positive and negative factors (*e.g.* Prison Population, Nonprison Population, Agriculture, Nonagriculture) are arranged as in the table. The result appears reasonable when it is noted that men usually engaged in agriculture formed only $690/3,000 = 0.23$ of the prison population, but $1,100,000/2,000,000 = 0.55$ of the nonprison population.

Notice that $Q = 0$ if $vw = ux$, or if $u/w = v/x$; that $Q = +1$ if v and/or w is 0; and that $Q = -1$ if u and/or x is 0. In other words, in Table 61, Q would show (1) zero association if the cell frequencies represented a purely random distribution of the table totals; (2) perfect positive association if all of the prison population, and/or none of the nonprison population was engaged in agriculture; (3) perfect negative association if none of the prison population, and/or all of the nonprison population was engaged in agriculture. The requirement for perfect association is less stringent than if "and/or" was replaced by "and" above, but Q is appropriate for treating the data of Table 61, if we are interested in the proportion of the prison population drawn from agriculture, as compared with the proportion of the nonprison population drawn from agriculture.

Should we want to measure the extent to which farmers and prisoners are strictly identical or exclusive categories, we may use the formula

$$r_4 = \frac{ux - vw}{\sqrt{_1n_2n.n_1n_2}} = \frac{\chi}{\sqrt{N}}, \qquad (108)$$

which assumes $v = w = 0$ (Table 61) for perfect positive association, $u = x = 0$ for perfect negative association, and (like Q) $vw = ux$ for no association. For Table 61,

$$r_4 = \frac{(690)(900,000) - (1,100,000)(2,310)}{\sqrt{(1,100,690)(902,310)(3,000)(2,000,000)}},$$

$$r_4 = -.025.$$

In view of the fact that the proportion of agriculturalists in the prison population was under half that in the nonprison population, the value of r_4 seems to be entirely too low, while the value of Q is about what would be expected. It seems extreme to insist that for perfect negative correlation the total nonagricultural population, but not a single farmer, must be in prison, as the formula for r_4 requires. For other problems, however, r_4 may be more appropriate than Yule's Q. This suggests that the choice of a measure of correlation should be adapted to the particular problem and interest of the investigator.

The two coefficients, Yule's Q and r_4, are both designed for the special case where the frequencies are impressionistically divided into two groups, or, in geometric terms, roughly collected at two discrete points. In Table 61, these points are Agriculture and Nonagriculture for one factor, and Prison population and Non-prison population for the other.

When the frequencies are distributed along two quantitative scales, and on each scale they are divided into two groups by a mark on the scale, and it is desired to find the amount of correlation between the paired scale values rather than between the proportions of cases in the two dichotomies, the so-called *tetrachoric* method is appropriate if the underlying mathematical assumptions mentioned below can be met. In Table 63, the factor, size of household, is reduced to two classes from the quantitative distribution on Table 62; the other factors, Relief and Nonrelief, is qualitative, like the categories of Table 61.

There are difficulties in the computation of the tetrachoric coefficient, r_t, but an approximate formula is:

$$r_t = \frac{1}{hk}\left\{1 - \frac{1}{N}\sqrt{N^2 - \frac{2hk(vw - ux)}{HK}}\right\}, \qquad (109)^1$$

where, reading from a table of normal areas and ordinates (Appendix Table 1),

h is the $\frac{x}{\sigma}$ value at $.5 - \frac{1^n}{N}$ or $.5 - \frac{2^n}{N}$

k is the $\frac{x}{\sigma}$ value at $.5 - \frac{n_1}{N}$ or $.5 - \frac{n_2}{N}$

H is the height of the ordinate at h,

K is the height of the ordinate at k,

and the other symbols have the same meanings as in Table 61.

The derivation of formula (109) assumes that both of the series (*e.g.*, size of households and relief-nonrelief) are normally distributed, that both dichotomies are continuous, that the

TABLE 62.—DISTRIBUTION OF RURAL RELIEF AND NONRELIEF HOUSEHOLDS BY SIZE, OCTOBER, 1933*

Size of household	Households		
	Relief	Nonrelief	Total
10 persons and over..................	290	246	536
9 persons...........................	202	213	415
8 persons...........................	353	336	689
7 persons...........................	493	560	1,053
6 persons...........................	633	997	1,630
5 persons...........................	834	1,322	2,156
4 persons...........................	846	2,061	2,907
3 persons...........................	846	2,408	3,254
2 persons...........................	745	2,430	3,175
1 person............................	358	627	985
All households......................	5,600	11,200	16,800

* Adapted from Thomas C. McCormick, Comparative Study of Rural Relief and Nonrelief Households, p. 88, *Research Monograph* II, Works Progress Administration, Division of Social Research, Washington, D. C., 1935.   Mid-point of last interval taken as 11.

[1] An alternative formula is

$$r_t = -\cos\left[\frac{\pi\sqrt{vw}}{(\sqrt{ux} + \sqrt{vw})}\right], \qquad (110)$$

where

$$\pi = 180°,$$

the symbols are arranged as in Table 63, and the sign of r_t is interpreted as in the case of Yule's Q above.

total frequency of the table is large, that the dichotomous divisions are not made too far toward the extremes of their distributions, and that the relationship is linear. If the table is not normal, the value of r_t is affected by the point of division of the dichotomies, *i.e.*, by whether each series is divided in the middle of the scale or at some other point.

In view of these restrictions, it hardly seems legitimate to apply r_t to Table 61 above. As Fig. 49 shows, the dichotomous
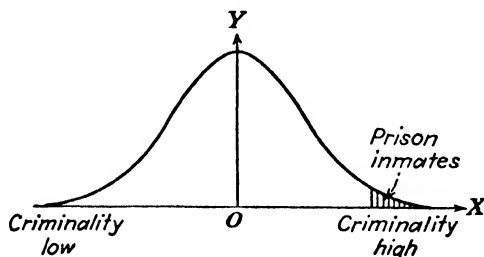
FIG. 49.—Proportion of adult male population in prison, Table 61.

line is drawn at the far upper end of the distribution of criminality, where the value of r_t is very sensitive to any skewness in the tail of the curve.

TABLE 63.—NUMBER OF RURAL RELIEF AND NONRELIEF HOUSEHOLDS CONTAINING LESS THAN FOUR PERSONS, AND FOUR PERSONS AND OVER, OCTOBER, 1933*

Size of household	Frequency		
	Relief	Nonrelief	Total
4 persons and more............	3,651 (u)	5,735 (v)	9,386 ($_1n$)
3 persons and less.............	1,949 (w)	5,465 (x)	7,414 ($_2n$)
All households................	5,600 (n_1)	11,200 (n_2)	16,800 (N)

* The data in the table should be so arranged that the value of the independent factor (size of household) increases from the bottom row to the top row, and the value of the dependent factor (economic independence) increases from the first column (relief) to the second column (nonrelief).

An inspection of Table 62 suggests that the distribution by size of household is somewhat skewed. We can test this, however, by shifting the position of the dichotomous line, and noting the effect on the value of r_t. If r_t remains rather stable, it is evidence that the distribution is normal enough for the use of the tetrachoric method. There is no way to judge the normality

of the dichotomous series, relief-nonrelief, except by rationalization. If we believe that the families in the two groups together would form an approximately normal distribution when classified according to some index of dependency, say, per capita family income, we may be justified in proceeding. The size of household series is continuous (*i.e.*, without gaps in the scale), and the same may be said of the relief-nonrelief series, although the nonrelief households were all nearest neighbors of the relief households, and represent a restricted nonrelief group. The dichotomous divisions in Table 63 do not fall too near the tails of their respective distributions. The condition that the total frequency should be large is also met. A rough test of linearity of relationship between the two series in the table could be made in the same way as was done in the case of biserial r (p. 200), but we shall not be risking a great deal if we dispense with it.

Turning to the definitions of h, k, etc., above, we find

$$\frac{1^n}{N} = \frac{9,386}{16,800} = 0.5587,$$

$$\frac{n_2}{N} = \frac{11,200}{16,800} = 0.6667,$$

so that

$$0.5 - 0.5587 = -0.0587,$$
$$0.5 - 0.6667 = -0.1667.$$

From Appendix Table 1 we read,
Corresponding to the "Area" entry 0.0587,

$$h = 0.14+.$$

Corresponding to the "Area" entry 0.1667,

$$k = 0.43+.$$

And the corresponding ordinates,

$$H = 0.395,$$
$$K = 0.364.$$

From Table 63 we have

$$N = 16,800,$$
$$v = 5,735,$$
$$w = 1,949,$$
$$u = 3,651,$$
$$x = 5,465.$$

Substituting in formula (109),

$$r_t = \frac{1}{(.14)(.43)} \left\{ 1 - \left(\frac{1}{16{,}800} \right) \right.$$
$$\left. \sqrt{(16{,}800)^2 - \frac{(2)(.14)(.43)[(1{,}949)(5{,}735) - (5{,}465)(3{,}651)]}{(.395)(.364)}} \right\},$$
$$r_t = \frac{1}{.0602} \left(1 - \frac{17{,}000}{16{,}800} \right),$$
$$r_t = -.21.$$

From Table 63, we see that 65 per cent of relief households have four or more persons, compared with only 51 per cent of nonrelief households. We therefore say that the degree of economic independence of a household is to a slight extent negatively correlated with the size of the household, as shown by the value $r_t = -.21$.

A quick method of finding the value of r_t is provided by L. Chesire, M. Saffir, and L. L. Thurstone's *Computing Diagrams for the Tetrachoric Correlation Coefficient*. We shall use one of these diagrams (Fig. 50) to test the normality of the size-of-household series in Table 62 by recomputing r_t after shifting the dichotomous line of division from three- to five-person households. The new groupings are shown in Table 64. The frequencies are reduced to proportions of the table total, 16,800, by multiplying the reciprocal of 16,800,

$$\frac{1}{16{,}800} = 0.00005952,$$

into each cell frequency. The proportions are entered in Table 65. We now take any row or column total that is not greater than .500 as a, any other column or row total at right angles to it as b,

TABLE 64.—NUMBER OF RURAL RELIEF AND NONRELIEF HOUSEHOLDS CONTAINING LESS THAN SIX PERSONS, AND SIX PERSONS AND OVER, OCTOBER, 1933

Size of household	Frequency		
	Relief	Nonrelief	Total
6 persons and more...................	1,971	2,352	4,323
5 persons and less...................	3,629	8,848	12,477
All households......................	5,600	11,200	16,800

TABLE 65.—FREQUENCIES OF TABLE 64 REDUCED TO PROPORTIONS OF THE
TABLE TOTAL, FOR USE WITH CHESIRE, SAFFIR, AND THURSTONE'S
COMPUTING DIAGRAMS

Size of household	Frequency		
	Relief	Nonrelief	Total
6 persons and more.............	−.117	+.140	.257
5 persons and less.............	+.216 = c	−.527	.743 = b
All households................	.333 = a	.667	1.000

and the proportion in the cell common to the a row (or column) and the b column (or row), as c. One set of these letters is indicated in the table. From Fig. 50, the diagram for $a = .33$,
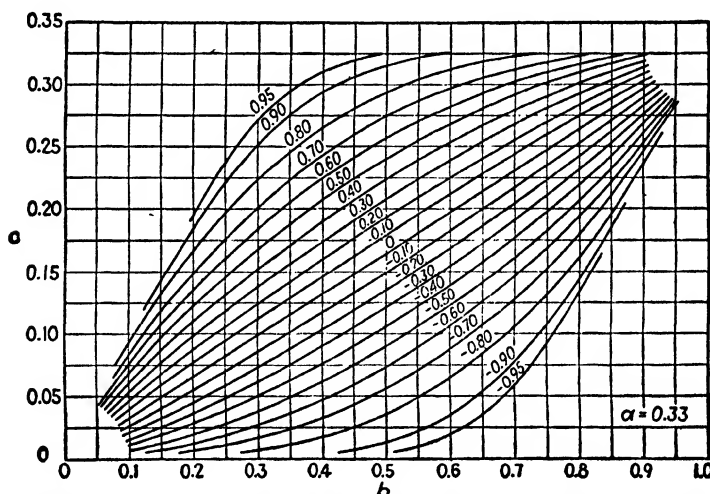
FIG. 50.—Sample computing diagram for the tetrachoric correlation coefficient. (From L. Chesire, M. Saffir, and L. L. Thurstone, *Computing Diagrams for the Tetrachoric Correlation Coefficient*, University of Chicago Bookstore, Chicago, 1933.)

we find at the intersection of the orthogonal[1] lines representing $b = .74$ and $c = .22$ a value $r_t = -.23$. If in Table 65 c falls in a positive quadrant, r_t has the sign shown in the diagram; but if c is in a negative quadrant, the sign indicated in the diagram is reversed. The signs of the quadrants are marked in Table 65, where it is seen that c is in a positive quadrant. Therefore,

[1] At right angles.

$r_t = -.23$, which agrees closely with the value of r_t computed for Table 63 with a different division of the dichotomy for size of households. So far as this test goes, then, the table seems to be normal enough to permit the use of the tetrachoric method. The test should be made for other points of division on the size-of-household scale, but would still be incomplete because new subdivisions cannot be tested in the relief-nonrelief series also.[1]

It should finally be observed that a fourfold correlation table includes some of the basic elements of experimental design. Thus, in Table 61, we have an independent factor or treatment, Agriculture; a dependent factor, Imprisonment for Crime; an experimental group, the Prison population; and a control group, the Nonprison population. On the other hand, dichotomies are used instead of classes based on measurement. In Table 61, sex and (roughly) age have been held constant, and there is nothing in the method that precludes as rigorous factor control as seems worth while. Even the broad 2×2 table may, therefore, be a valuable analytical device.

[1] If it is needed to determine the value of the tetrachoric coefficient, r_t, very precisely, the complete formula may be seen in several texts, *e.g.*, Davenport and Ekas, *Statistical Methods in Biology, Medicine and Psychology*, 4th ed., pp. 105–106, or Peters and Van Voorhis, *Statistical Procedures and Their Mathematical Bases*, p. 370; and helpful tables with explanations are given by Karl Pearson, *Tables for Statisticians and Biometricians*, 3d ed., Part I, pp. xxxvi, xliii, l, liii, 31, 32, 33, 34, 42–52, 52–57; Part II, pp. xliv, 73, 74. Formulas have been derived for the standard errors (see Chap. XII) of biserial r, the coefficient of tetrachoric correlation, and the coefficient of contingency. The standard error of the coefficient of contingency, C, is hardly needed if the value of χ^2 for the contingency table is referred to a table of χ^2, as was done above in the section on this coefficient. The formula may be seen, however, in such texts as Holzinger, *Statistical Methods for Students in Education*, p. 278. The standard error of the tetrachoric correlation coefficient, r_t, is also given by Davenport and Ekas, *op. cit.*, p. 108, and by Peters and Van Voorhis, *op. cit.*, p. 371. G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, p. 408, show the formula for the standard error of r_t. Somewhat simpler are the standard error formulas of r_{bis} and Q:

$$\sigma_{r_{bis}} = \frac{\sqrt{\dfrac{pq}{y} - r^2_{bis}}}{\sqrt{N}} \tag{111}$$

$$\sigma_Q = \frac{1 - Q^2}{2} \sqrt{\frac{1}{w} + \frac{1}{x} + \frac{1}{u} + \frac{1}{v}} \tag{112}$$

Exercises

1. What are the chief disadvantages in correlating qualitative series, as compared with quantitative series?

2. What preliminary test should be made of a qualitative table before applying correlation to it?

3. What is the amount of correlation between type of college training and success in teaching in the following table?

Two Hundred High School Teachers Classified by Type of College from Which They Graduated, and by Success in Teaching

Institution	Successful	Unsuccessful	Total
Teachers college......................	58	42	100
University or college..................	49	51	100
Total.............................	107	93	200

Defend your choice of a coefficient, and explain the meaning of your results.

4. How much association, if any, is there between the sex of distinguished people and the socioeconomic class of their fathers in the table below?

Famous British Men and Women Classified by Social Origin*

Socioeconomic class of father	Men	Women
Nobleman............................	1,059	108
Gentleman...........................	724	83
Politician, lawyer.....................	666	61
Soldier, sailor........................	490	53
Divine...............................	1,100	57
Teacher.............................	274	23
Physician............................	396	35
Administrator........................	194	12
Writer, artist........................	371	109
Businessman.........................	929	95
Artisan..............................	446	38
Laborer, servant.....................	81	8
Agriculture..........................	270	18
Total.............................	7,000	700

* Adapted from Table 4, p. 708, Joseph Schneider, Class Origin and Fame: Eminent English Women, *American Sociological Review*, Vol. 5, pp. 700–713, 1940.

Is the association positive or negative? What does the association mean in terms of this problem? What should be done about the correction for broad grouping in this case? Is the value of C significantly greater than zero? Explain what this means.

5. What is the amount of association between the sex of a sample of undergraduate students at the University of Wisconsin in 1938–1939 and their state of residence? What coefficient is most appropriate to this problem, and why? Interpret its meaning.

A SAMPLE OF UNDERGRADUATE STUDENTS, UNIVERSITY OF WISCONSIN, 1938–1939, CLASSIFIED BY SEX AND BY STATE OF RESIDENCE

State of residence	Male	Female	Total
Wisconsin...............................	94	44	138
Other...................................	17	27	44
Total.................................	111	71	182

6. Find the amount of association between type of offense and body build in the table:

CRIMINALS CLASSIFIED BY TYPE OF OFFENSE AND BODY BUILD*

Body build	First-degree murder	Second-degree murder	Assault	Robbery	Burglary and larceny	Forgery and fraud	Rape	Other sex	Vs. public welfare	Arson and all other	Total
Slender....	42	79	7	54	213	57	18	18	31	7	526
Medium...	155	358	49	244	1004	260	119	80	127	71	2467
Heavy.....	77	147	18	81	302	110	44	46	77	15	917
Total....	274	584	74	379	1519	427	181	144	235	93	3910

* From A. E. HOOTON, *The American Criminal*, Vol. I, Appendix, IX-8, Harvard University Press, Cambridge, 1939.

Is the value of the coefficient significantly greater than zero? What does the coefficient mean here?

7. What is the amount of correlation between the age distributions of females in the neighboring urban and rural counties in the accompanying table of age distributions?

AGE DISTRIBUTIONS OF FEMALES IN A RURAL (RUTHERFORD) AND A NEAR-BY
URBAN (MECKLENBURG) COUNTY IN NORTH CAROLINA, 1930*

Age, years	Rural county	Urban county
Under 5..................	2,553	6,542
5–9......................	2,846	7,311
10–14....................	2,428	6,424
15–19....................	2,247	6,751
20–24....................	2,109	7,862
25–29....................	1,579	6,990
·30–34....................	1,201	5,277
35–44....................	2,202	8,288
45–54....................	1,520	5,199
55–64....................	932	2,548
65–74....................	515	1,342
75 and over..............	237	586
Total..................	20,369	65,120

* From Fifteenth Census of the United States, 1930, Bureau of the Census.

References

DAVENPORT, C. B., and M. P. EKAS: *Statistical Methods in Biology, Medicine, and Psychology,* 4th ed., pp. 97–108, John Wiley & Sons, Inc., New York, 1936.

ELDERTON, W. P.: *Frequency .Curves and Correlation,* 3d ed., Chap. IX, University Press (John Wilson & Son, Inc.), Cambridge, Mass., 1938.

HOLZINGER, KARL J.: *Statistical Methods for Students in Education,* pp. 271–272, Ginn and Company, Boston, 1928.

PETERS, C. C., and W. R. VAN VOORHIS: *Statistical Procedures and Their Mathematical Bases,* Chaps. XIII and XIV, McGraw-Hill Book Company, Inc., New York, 1940.

YULE, G. U., and M. G. KENDALL: *An Introduction to the Theory of Statistics,* Chaps. III and V, pp. 252–253, 408, 410, Charles Griffin & Company, Ltd., London, 1937.

CHAPTER XII

SAMPLING AND SAMPLING ERRORS

1. Definitions.—In sociological research, it is seldom possible to study more than a part of the whole, or *universe*,[1] in which we are interested. For example, if it is wanted to know whether the educated or the uneducated in the United States have the higher birth rate, it would be impractical to find the birth rate of the millions in each class. A sample would have to be taken of each group, and the birth rates of the two samples compared. If the samples were large and properly taken, the sample birth rates should be rather close to the true rates for the total educated and uneducated in the country.

A value (*e.g.*, a mean) found from a sample is called a *statistic*, whereas the corresponding true or *expected* value in the universe is called a *parameter*. The primary purpose of all sampling is to learn something about a universe, often to estimate the value of a parameter from the value of a statistic. There is seldom any interest in a sample or in the value of a statistic for its own sake. A good sample is, therefore, one that yields reliable information about a universe.

The first step in sampling is to define the universe to be sampled. Thus we might define the universe of the educated as consisting of all married couples in the United States living together through the year 1939 who had successfully passed at least the first year of high school; and the universe of the uneducated as corresponding couples who had less schooling than this, the birth rates to be compared as of the year 1939. The sociological universe should usually be defined in both space and time.

Since the universe is made up of events, a definition of the *event* is also necessary. In our illustration above, the event is a married couple with a birth or a married couple without a birth during 1939. There are thus two kinds of events, couples with

[1] Synonymous terms often used are *population* and *parent*.

221

a birth, which may be called *successes*, and couples without a birth, which may be called *failures*. The word "success" merely designates that particular event among two or more different kinds of events in which the investigator is chiefly interested. If we were sampling farmers to find their net annual incomes in 1939, the event would be a farmer's income, and would represent a continuous, measured variable. In the case of a measured variable, there is, of course, no dichotomy of success and failure, but merely a number of specific values.

A universe may consist of an *infinite* or of a finite (*limited*) number of events. If the number of events is very large, the universe may be regarded as infinite for practical purposes.

The events in a universe may already have happened, or may be yet to happen. In the former case they are said to be *existent;* in the latter, *hypothetical*. In our illustration above, at the beginning of the year 1939 none of the events (a birth or the absence of a birth to a married couple) has happened; at the end of the year, all of them have happened. Similarly, heads or tails is a hypothetical event before tossing a penny, an existent event after tossing. When the universe to be sampled consists entirely of completed events, the universe is said to be existent; when it consists entirely or partly of events yet to come, it is said to be hypothetical. Prediction, with which social science must be concerned, is of course possible only with respect to hypothetical universes, since we do not "predict" past events.

It is also important to notice whether the universe to be sampled is to be regarded as a *unique*, historical set of events (situation), as a constant or *recurrent* situation or system of causes, or as a *changing* situation. If we are interested in the death rate from the influenza epidemic of 1918, we have a unique universe. But if we attempt to predict the rate of mortality in Chicago, we assume a continuous or recurrent, *i.e.*, essentially unchanging, universe. As a matter of fact, strictly continuous or recurrent universes never occur in social research, since there is constant change in the complex of factors that compose any social situation. The important question, therefore, is whether the universe can be expected to be approximately recurrent, or unchanged, over a period in which we are interested. If so, we may be justified in trying to predict what will happen in that

period on the strength of what has occurred. It is sometimes possible to discover the nature, direction, and rate of change in a changing universe, so that we can allow for it in making a prediction.

Finally, we shall find it worth while to distinguish between *homogeneous* and *heterogeneous* hypothetical universes. A universe is homogeneous when each hypothetical event has the same a priori probability of becoming a success or a specified value of a variable; it is heterogeneous when this probability is not the same for each hypothetical event. A homogeneous universe derives from a single set or system of causes, a heterogeneous universe from two or more distinct sets of causes, as judged by their effects on the hypothetical events in which we are interested. When an insurance company sets up a class of "risks," composed of, say, males, native white, married, in the legal profession, aged 25, class "A" medical examination, living in Michigan, the company is trying to create a homogeneous universe. Every person or hypothetical event admitted to the class must be judged alike in respect to certain characteristics that are believed to be related to the event, death. In other words, each member of the class must have the same apparent chance of death. In this way, and by requiring that the conditions of life for the class must go on essentially in the future as in the past (*e.g.*, in case one of the insured persons enlists in a war, his contract may be modified or canceled), some likelihood is created that the system of causes affecting the mortality rate of the class will continue each year about the same as the year before, except for chance factors. If, however, a number of men aged 65 were to be admitted to the risk class originally composed of men aged 25, heterogeneity in the hypothetical events would at once be introduced. While such a mixed or heterogeneous universe might be recurrent if the proportion of the two ages were kept constant, it could no longer claim to be homogeneous, because the chance of death is known to be different for a man aged 25 and a man aged 65. In practice, just when a hypothetical universe may be considered homogeneous is a matter of information and of degree. Of course, no two persons in a life-table category actually have exactly the same chance of death. The more completely the causes that are related to the success are controlled and equated from event

to event, however, the more accurate and reliable the prediction from the sample will tend to be, within limits. Where to stop the effort to increase homogeneity is a question of judgment and expediency. The more homogeneous the categories of any classification are made, the greater their number, and the fewer the events that will fall in any one category. While it is usually advisable to sacrifice the size of the sample for the sake of homogeneity to a certain point, diminishing returns set in if the idea is carried too far.

2. Taking the Sample.—The events in a sample may be drawn from the universe (1) at random, (2) at regular intervals, (3) at random from different strata or subclasses of the universe, or (4) according to some purposive scheme, such as from the middle and ends of a distribution. Thus, (1) we might draw marriage certificates at random from an alphabetical list of all of the marriage certificates in a file, (2) we might take every fifth certificate in order, (3) we might draw a proportional number of certificates at random from each separate county and city list, or (4) we might take certificates from the top, middle, and bottom of the list. The most common method of taking a sample, and the one to which most of the statistical theory of sampling applies, is the random. The method of sampling at random proportionally from within strata—*e.g.*, marriage certificates taken at random from each county file—is more representative than random sampling from the total universe—*e.g.*, certificates taken from a grand list. Unfortunately, however, the sampling errors of only a few statistics are available in the case of *stratified sampling*. Purposive sampling is seldom as reliable as either of the other two methods, and difficulties of determining sampling errors are encountered. We shall deal here primarily with random sampling, but shall introduce stratified sampling for a mean and for a proportion.

A sample is *random* when at any given draw or trial, considered alone, every existent event has an equal chance of being taken, or every hypothetical event is equally likely to occur. In other words, in a random sample the chance of being "drawn" or "thrown" is independent of the character of the event. In addition, a *simple* sample—also called a *Bernoulli* sample, after a French mathematician who studied it—requires that the probability of drawing or throwing a success or a specified value

shall remain the same from one draw or trial to another.[1] It is theoretically possible to random sample any universe, but a simple sample can be drawn only from an infinite universe. Suppose we have an existent universe of 1,000 marriage certificates and wish to take a random sample of 100. Suppose, further, that 60 of these marriages have ended in divorce. At the first draw, the probability of taking a marriage that has ended in divorce is $\frac{60}{1000}$, or 0.060. If we happen to draw a divorced marriage at the first trial, the probability of getting a divorced marriage at the second draw will be $\frac{59}{999}$, or 0.059; otherwise it will be $\frac{60}{999}$, or 0.06006. Now if the certificates are drawn entirely independently of the question of divorce, at any draw one certificate will have as good a chance of being taken as any other in the file, and the resulting sample will be random. But because the probability of drawing a given event, say, a divorced marriage, changes from one draw to the next in the limited universe of 1,000 certificates, the sample will not be simple. If, however, after each draw of a certificate from the file of 1,000, its number is recorded in the sample and the certificate is returned to the file, the number of certificates in the file will remain constant and the probability of drawing a divorced marriage will not change from draw to draw. By the act of replacement the universe becomes infinite. Of course, if we happen to draw the same certificate more than once, it will have to be accepted in the sample each time it is drawn, if it is wanted to maintain an infinite universe.

In the case of a hypothetical universe, a simple sample of hypothetical events can be drawn only if the universe is homogeneous, like a life insurance risk group. All the causes that determine the chance of death that the actuaries have been able to consider must be the same for each individual in the group. If a random sample were drawn from a mixture of two different risk groups, so that, say, persons of different sexes were included in the sample, the chance of death would not be the same from one hypothetical event to another (person to person), and the sample would be further removed from a simple sample.

[1] But it is not assumed that the probability is the same for *different kinds* of events or different values, *e.g.*, a person of age 25 and a person of age 30 in a universe of ages.

It follows from the preceding definitions that a simple or random sample of an existent universe is not necessarily a simple or random sample of the hypothetical universe from which the existent universe was derived.   But when the existent universe itself is a simple or random sample of a hypothetical universe, a simple random sample of the existent universe will be a simple or random sample of that hypothetical universe also.

In practice, it is not easy to obtain a random sample.   The most manageable case is that of a limited existent universe each of whose events can be individually identified, such as the list of marriage certificates mentioned above.   If we take certificates or pages of certificates at regular intervals from an alphabetical or other random list, say every twentieth certificate or page, the first page being chosen at random, the sample should apparently be random, because there is no obvious connection between this order and the information on the marriage certificates.   If the interval is not too large, this method should also be more representative than other types of random sampling, since it takes certificates proportionately from every part of the list.   There are many other devices for taking a random sample from a list.   One of the commonest is to number the items, place corresponding numbers on tickets in a box, shuffle them, and draw.   Experience has shown, however, that methods like the above will not always yield a random sample.   Mathematically, the ideal plan is to draw the sample from a table of random numbers, such as L. H. C. Tippett's *Random Sampling Numbers*,[1] which are combinations of digits taken at random from census reports.   A specimen page of these numbers is shown below (Fig. 51).

Imagine that we wish to take a random sample of 200 marriage certificates from a list of certificates in a state file.   The certificates are filed and numbered consecutively, so that any *n*th certificate from the beginning of the list can be quickly located. The smallest number printed in the table is 0,000 and the largest is 9,999.   If the number of events in the universe is close to 10,000, we can simply go down each column of four figures in the table, taking for our sample the first 200 numbers within the range of our universe that we meet.   When the universe is

1 2 3 4	5 6 7 8	9 10 11 12	13 14 15 16	17 18 19 20	21 22 23 24	25 26 27 28	29 30 31 32
2 9 5 2	6 6 4 1	3 9 9 2	9 7 9 2	7 9 7 9	5 9 1 1	3 1 7 0	5 6 2 4
4 1 6 7	9 5 2 4	1 5 4 5	1 3 9 6	7 2 0 3	5 3 5 6	1 3 0 0	2 6 9 3
2 7 3 0	7 4 8 3	3 4 0 8	2 7 6 2	3 5 6 3	1 0 8 9	6 9 1 3	7 6 9 1
0 5 6 0	5 2 4 6	1 1 1 2	6 1 0 7	6 0 0 8	8 1 2 6	4 2 3 3	8 7 7 6
2 7 5 4	9 1 4 3	1 4 0 5	9 0 2 5	7 0 0 2	6 1 1 1	8 8 1 6	6 4 4 6
5 8 7 0	2 8 5 9	4 9 8 8	1 6 5 8	2 9 2 2	6 1 6 6	6 0 6 9	2 7 6 3
9 2 6 3	2 4 6 6	3 3 9 8	5 4 4 0	8 7 3 8	6 0 2 8	5 0 4 8	2 6 8 3
2 0 0 2	7 8 4 0	1 6 9 0	7 5 0 5	0 4 2 3	8 4 3 0	8 7 5 9	7 1 0 8
9 5 6 8	2 8 3 5	9 4 2 7	3 6 6 8	2 5 9 6	8 8 2 0	1 9 5 5	6 5 1 5
8 2 4 3	1 5 7 9	1 9 3 0	5 0 2 6	3 4 2 6	7 0 8 8	3 9 9 1	7 1 5 1
5 6 6 7	3 5 1 3	9 2 7 0	6 2 9 8	6 3 9 6	7 3 0 6	7 8 9 8	7 8 4 2
1 0 1 8	6 8 9 1	1 2 1 2	6 5 6 3	2 2 0 1	5 0 1 3	0 7 3 0	2 4 0 5
6 8 4 1	5 1 1 1	5 6 8 8	3 7 7 7	7 3 5 4	3 4 3 4	8 3 3 6	6 4 2 4
2 0 4 1	2 2 0 7	4 8 8 9	7 3 4 6	2 8 6 5	1 5 5 0	5 9 6 0	5 4 7 9
5 5 6 5	4 7 6 4	2 6 1 7	5 2 8 1	1 8 7 0	6 4 9 7	5 7 4 4	9 5 7 6
4 5 0 8	1 8 0 8	3 2 8 9	3 9 9 3	9 4 8 5	4 2 4 0	2 8 3 5	9 9 5 5
2 1 5 2	6 4 7 3	5 6 9 2	9 3 0 9	7 6 6 1	1 6 6 8	5 4 3 1	7 6 5 8
6 9 1 7	4 1 1 3	7 3 4 0	6 8 5 3	1 1 7 2	7 2 2 9	1 2 7 9	5 0 8 5
8 2 4 1	4 1 2 4	4 1 3 1	9 5 0 0	5 6 5 7	3 9 3 2	5 9 4 2	3 3 1 7
7 9 1 3	3 7 0 9	5 9 4 4	9 7 6 3	2 7 5 5	4 2 1 1	4 9 9 5	8 6 5 7
9 3 8 5	7 1 2 5	3 2 3 0	0 7 3 7	2 9 5 7	1 0 1 3	6 3 6 9	4 4 9 4
3 4 3 6	6 2 9 3	6 0 2 5	9 3 8 4	3 3 4 3	1 0 7 1	1 4 6 8	4 8 0 1
9 0 9 4	1 6 3 4	5 0 7 0	0 6 6 4	6 5 1 0	0 9 1 8	4 6 0 1	4 2 9 4
9 2 2 6	9 2 9 6	2 7 9 6	7 0 9 7	4 0 5 7	2 0 7 4	6 2 9 7	2 5 8 7
7 7 8 1	3 7 6 0	2 8 9 5	7 6 5 3	0 0 9 1	7 0 1 2	1 3 0 8	1 0 9 6
9 7 4 2	9 6 9 4	7 3 4 7	0 0 1 7	9 5 7 2	1 8 5 0	0 1 1 6	1 8 9 9
9 4 2 0	9 2 1 0	8 7 8 7	9 3 7 5	4 6 6 3	0 3 9 6	6 7 1 7	5 5 6 2
1 1 7 9	3 5 7 1	5 9 9 2	3 0 5 9	9 0 1 5	5 6 0 8	2 3 4 8	8 1 4 4
0 7 0 8	4 0 1 1	4 0 5 7	1 5 5 0	1 6 7 4	1 3 7 6	5 2 4 3	4 4 2 7
6 3 5 0	3 9 9 6	3 7 9 5	2 1 7 6	8 1 8 2	4 5 1 4	6 3 4 9	3 4 8 3
1 4 1 4	7 1 5 2	3 6 5 8	1 6 3 5	0 6 3 8	3 4 4 3	4 4 4 0	3 0 8 6
7 0 4 1	8 9 8 5	7 0 1 1	5 6 7 6	7 5 7 0	6 6 8 5	1 7 7 6	3 1 5 4
3 2 4 3	2 7 8 3	0 8 4 0	9 0 5 4	8 8 6 2	5 1 7 3	8 4 3 3	9 1 1 7
7 9 2 2	4 9 3 1	5 7 5 3	6 1 6 0	6 5 6 6	8 6 0 2	3 4 2 3	9 0 7 4
8 7 6 9	3 5 1 3	8 9 7 6	0 7 8 0	6 3 8 2	0 0 2 9	2 6 1 9	5 9 8 2
2 5 1 0	7 2 7 4	8 7 4 3	0 0 0 0	1 8 5 0	2 4 0 8	3 6 0 2	5 1 7 9
0 2 2 4	2 4 0 4	9 8 1 1	6 6 4 1	9 7 3 2	1 6 6 2	9 1 5 8	1 4 0 4
3 0 0 9	8 5 1 6	7 2 4 5	9 4 0 9	2 8 4 4	0 7 1 7	1 0 7 2	3 1 3 7
7 4 8 9	0 2 2 1	7 9 2 1	2 3 5 1	2 6 9 6	4 9 0 6	2 4 8 4	3 8 6 8
5 1 8 8	1 8 2 5	2 2 2 0	9 3 8 2	0 5 3 2	1 9 1 5	1 7 9 0	2 0 8 1
1 1 9 8	2 5 4 5	2 4 8 2	9 6 0 7	0 0 6 7	3 7 4 4	9 8 6 6	5 0 9 6
3 9 0 8	4 6 7 6	7 8 1 6	6 5 1 7	9 1 2 1	3 1 7 1	4 1 1 9	3 6 1 5
1 0 9 4	2 2 2 3	1 6 7 5	2 2 8 2	3 7 1 2	8 1 9 1	1 3 3 0	1 4 5 4
1 8 1 7	7 7 2 3	5 5 8 2	7 1 5 3	9 5 1 8	0 2 3 1	7 7 8 2	5 7 4 2
6 2 0 8	9 5 9 8	9 6 2 3	2 1 1 4	7 7 4 7	2 0 9 6	5 0 2 7	0 5 6 1
1 7 5 2	4 5 1 9	2 7 4 9	8 0 2 0	4 6 4 2	1 1 9 0	7 3 0 2	8 3 5 0
0 4 8 6	6 9 9 3	3 1 1 5	5 0 2 5	4 8 8 7	1 5 7 1	9 8 1 9	6 8 0 4
1 9 4 2	3 0 0 4	1 4 4 2	2 8 1 0	1 4 7 9	0 9 7 0	7 3 0 2	3 7 7 5
4 9 3 0	9 7 8 5	7 4 6 0	3 9 9 6	2 8 6 4	0 5 5 9	3 9 8 5	8 0 9 2
2 3 4 9	1 5 9 4	7 1 5 2	0 2 5 7	4 0 4 1	4 1 0 5	3 1 8 0	9 8 0 6

FIG. 51.—Specimen page of random sample numbers. (*From Tracts for Computers, Number XV, Random Sampling Numbers, ed. by Karl Pearson, arranged by L. H. C. Tippett, p. 1.*)

much smaller than 10,000, it sometimes saves time to assign several table numbers to each event. For example, if there are only 2,000 events (*e.g.*, marriage certificates) in the universe, we may assign to event number 1 the table numbers 0 through 4, to event number 2 the table numbers 5 through 9, and so on. If then we read, say, the number 0061 from the table, we draw event number 13 from the universe ($\frac{60}{5} + 1 = 13$).[1] The same event is accepted only once from a limited universe, regardless of how many times it may be drawn. Also, the number of digits to read in the table may correspond to the number of digits needed to express the total of events in the universe. Thus, if the universe contains 800 events, we may draw three-digit numbers, *e.g.*, 295, 016, 273, and so on; if the number of events is 500,000, we may draw six-digit numbers, such as 295,266; 416,795; 003,074. The table may be read in any direction or order.

When the individual events of a limited universe cannot be identified and labeled, probably the next best thing is to identify *groups* of them, usually on a geographical-time basis. Thus a random sample of the farmers of a state, of whom there is no list, may be obtained by numbering each township in the state and drawing a random sample of townships by one of the methods suggested above. Then each township drawn in the sample may be visited at a given date, and all the farmers in it taken in the sample. Or, if necessary, the sample townships may be divided into school districts and a random sample of these districts drawn before going to the events (farmers) themselves. When this approach has to be used, the number of groups composing the universe should be as large as practicable, while the number of events in each group should be a minimum and nearly equal from group to group. For example, if the township is the smallest unit for which data are available, a township with a large population may be subdivided and represented by two or three tickets, instead of by one, in the drawing, so that the probability of drawing a township will be roughly proportional to the size of its population.

In dealing with an infinite or a very large universe, it is of course not possible to list and label all the individual events, but

---

[1] In this case, to find the serial number, if the table number is not already an exact multiple of five, reduce it to the nearest multiple of five, divide by five, and add one.

it may be feasible to use the group method mentioned in the preceding paragraph. For example, if a physical anthropologist wanted to sample the white race, he might divide the countries occupied by the various branches of this race into small geographical areas, number them, and draw them at random. He would then probably have to go to each of the areas drawn, further subdivide them, draw a random sample of the small subdivisions, and then finally perhaps take a random sample of the individuals living in each subdivision.

Such a plan as the above, however, is not adapted to a hypothetical universe, like the number of heads or tails that might be thrown with a penny, or the number of divorces that might occur in the United States over some future period of time. The only way to draw a random sample in this case is to define a set of conditions, or causal system (*e.g.*, social conditions in the United States, Jan. 1, 1940 to Jan. 1, 1941), draw at random a number of hypothetical events that satisfy the conditions (couples married on Jan. 1, 1940), and let the system act to convert them to existent events (couples divorced, not divorced on Jan. 1, 1941); or else wait until the system has produced a large number of existent events (couples married Jan. 1, 1940, *after* Jan. 1, 1941), and then draw at random as many of them as are needed for the sample. In either case, if a simple sample is wanted, it is, of course, necessary to make sure that the existent events (couples divorced or not divorced Jan. 1, 1941) were derived from hypothetical events (couples married Jan. 1, 1940), each of which had (on Jan. 1, 1940) essentially the same a priori probability of becoming a success (divorced couple) throughout the experiment (Jan. 1, 1940 to Jan. 1, 1941), except for chance factors. Notice also that a causal system that does not act uniformly over its time cycle must furnish sample events from the whole of its cycle, to avoid important omissions. For example, in determining the death rate of infants during the first year of life, observations should extend over the complete period of 12 months, because the death rate is subject to seasonal variation.

If a heterogeneous hypothetical universe, *i.e.*, a hypothetical universe in which the chance of success is not the same from one hypothetical event to another (*e.g.*, a class of life insurance risks of different ages, where p is the probability of an individual's death within the year, and a hypothetical event is a person taken

at the beginning of the year), exists without important change over a period of time, then a random sample drawn from a large number of the events at the end of this period will yield an estimate of the *mean probability* of death for the mixed class.

When a universe is divided into strata with respect to some trait, a proportional[1] simple subsample is taken from each stratum, and these subsamples are combined, the resulting total sample is called a *Poisson* sample, in honor of the French mathematician who described it. Thus, for the purpose of drawing a Poisson sample, an existent universe of family incomes in New York City in 1939 may be divided into the classes: Under $500, $500–$999, $1,000–$1,499, . . . ; or, supposing that we are interested in divorce, we may define a hypothetical universe of ever-married women in the city of Philadelphia on Jan. 1, 1940, consisting of subgroups whose members are alike in respect to occupation of husband, presence or absence of children, religious affiliation, length of time since marriage, and so on. If all the requirements of Poisson sampling are to be met, each stratum must constitute an infinite subuniverse. In the case of our existent universe of family incomes, this will be approximately true if the number of incomes in each class is very large. Any hypothetical universe or stratum may be regarded as infinite on the assumption that the defined set of conditions theoretically acts to produce events without limit. For example, it may be reasoned that the conditions that produced a certain percentage of divorces among a group of Philadelphia women whose husbands were skilled laborers, who had borne children, who were Protestants, who had been married five years, and so on, might continue indefinitely to produce the same percentage of divorces (except for random errors) among women of this description. As a rule, however, it is more realistic to consider how long we may expect a hypothetical universe actually to persist without important change, and then decide whether the probable number of events that will be produced in a given stratum within that period may be regarded as infinite for practical purposes. To refer again to our illustration, we might conclude that the set of conditions responsible for the divorce rate observed in the class of women defined above would probably remain essentially

[1] Preferably also weighted by the value of the standard deviation of the stratum or subgroup.

the same no longer than perhaps a decade, but that in 10 years several thousand women would come within the class, a number great enough to be regarded as infinite without noticeable error.

If a simple sample is drawn from only one of several strata forming a universe, and from it an attempt is made to judge the whole universe, the sample is called a *Lexis* sample, after a German statistician. The sampling error of a Poisson sample is less than that of a random sample, while that of a Lexis sample is greater. A Lexis sample is seldom taken intentionally, but may occur when some important part of the universe is omitted from the sample.[1] The Poisson sample, on the other hand, is the most representative sample that can be taken of sociological data, and should be used much more than it now is.

What has been said above about the sampling of an *attribute* (an unmeasured quality called an event, such as the survival or death of an insured person) applies equally to the sampling of a *variable* (a measured quality, such as the net annual income of a farm family). In the case of sampling a variable, the parameter in which we are usually interested is the mean of the values in the universe (*e.g.*, the mean net annual income of the farmers in Nebraska), although it may be the standard deviation, a correlation coefficient, or other index.

When the purpose in taking a sample is to use the proportion, mean, or other statistic from the sample as an estimate of the corresponding parameter in the universe, it is needed to know the range of error in the estimate due to sampling. This can be found only if the sample is approximately of some standard type, such as random, simple, or Poisson. Thus, if we find from a simple sample of juvenile delinquents that 21 per cent were from broken homes, we are able to estimate with the aid of the mathematical theory of sampling that the chances are, say, 19 to 1 that certain limits, say 15 to 27, will enclose the true percentage from broken homes in the universe from which the sample was taken. If the nature oˆ the sample is uncertain, so that we do not know that it is, say, simple or Poisson, we cannot apply the appropriate formulas for finding the errors of sampling, and so cannot gauge the amount of error in any statistic estimated from

[1] See BRUCE D. MUDGETT and S. R. GEVORKIANTZ, Reliability of Forest Surveys, *Journal of the American Statistical Association*, Vol. 29, pp. 257–281, 1934.

the sample. The chief assurance that we can have about the nature of a sample must come from a knowledge of the method by which it was taken. Thus, we must know that the conditions of, say, simple sampling were at least broadly met in drawing the sample before we can safely treat it as a simple sample.

3. The General Theory of Sampling.—In general, the theory of sampling that provides a basis for the measurement of sampling error is as follows. Suppose that we draw a large number, N, of random samples of equal size, n, from a universe of juvenile delinquents, and list the number of delinquents with broken homes (successes) in each sample. We shall then have a table like Table 66.

This is a *sampling distribution* of the number or *frequency* of successes per sample. We may find its standard deviation by the familiar formula,

$$\sigma = \sqrt{\frac{\Sigma f(X - M_x)^2}{N}},$$

where X is the number of successes per sample, M_x is the mean number of successes per sample, f is the number of samples having a given number of successes, and N is the total number of samples. Since this is the standard deviation of the number of successes from many actual samples, we may call it an *empirical standard error*, to distinguish it from the standard deviation of a series where the question of sampling does not enter. We may further call the standard error of this formula the *empirical standard error of the number of successes per sample*, to differentiate it from the standard error of, say, a mean or correlation coefficient.

An empirical standard error like the above has the disadvantage, however, that it is itself a sample value that is affected by the number of samples taken, and varies because of random errors of sampling. Mathematicians are able to calculate a more exact or *theoretical standard error*,[1] provided they are allowed to specify the nature of the distribution of the universe values and the conditions under which the sample is taken. This enables them to lay down requirements which ensure that the parameter,

---

[1] Or probable error, if preferred. From Chap. IX. pp. 160 and 161, it will be recalled that the probable error is related to the standard error by the equation $P.E. = .6745\sigma$, where $\sigma = \epsilon$ in our subsequent notation.

say, a frequency, will be distributed in the samples according to some established mathematical principle, such as the binomial theorem or the normal curve.

TABLE 66.—DISTRIBUTION OF BROKEN HOMES PER SAMPLE OF 50 JUVENILE DELINQUENTS IN 100 RANDOM SAMPLES

Broken Homes per Sample	Samples
0	0
1	0
2	0
.	.
.	.
8	1
9	1
10	2
11	5
12	9
13	10
14	12
15	12
16	11
17	8
18	9
19	5
20	6
21	4
22	3
23	1
24	0
25	1
26	0
27	0
.	.
.	.
48	0
49	0
50	0
Total..................	100

Some of the commonest of the standard error formulas that are applied in the sampling of attributes assume that the sampling distribution is *binomial* in type. It will be recalled[1] that N times

[1] See Chap. IX.

the binomial expansion shows how many of N random samples of n events each may be expected to have given numbers of successes from 0 to n, where the probability of success remains the same from event to event. If it can be shown that these requirements were at least approximately complied with in drawing the events of a sample, we may assume that the sampling distribution will be approximately binomial in form. The standard deviation of the binomial is well known, and is then the theoretical standard error that we are seeking:

$$\sigma_f = \sqrt{npq},$$

where p is the constant chance of success in the binomial universe, $q = 1 - p$, n is the number of events in a single sample, and f is the frequency. We have only to substitute in this formula to get the standard deviation (called standard error) of the sampling distribution.

In the investigation of sociological attributes, however, there is usually available only one sample, rather than a distribution of many samples. In that case, if the sample was taken under binomial conditions, and its size is large, the best estimate of p, the proportion of successes in the universe, is the proportion of successes in the single sample. This estimate of p is then used in the above formula to compute an approximate theoretical sampling error.

It will be noticed that the binomial theorem merely repeats the requirements of the simple sampling of attributes, which we have seen can be met only if an existent universe is infinite and well mixed, or a hypothetical universe is homogeneous. Because of the difficulties in taking a simple sample under many conditions in sociological research, it is fortunate that the standard errors of simple samples are usually not very different from those of random samples, and in any case are somewhat larger. For these reasons, investigators often apply simple sampling errors to random or even stratified samples, in order to save labor or to be on the conservative side when in doubt as to what the error formula should be.

4. Only Large Samples Considered.—In sociological investigations, many factors that the sociologist is unable to control usually cause small samples to differ radically from one another. Small samples are, therefore, not often used in social research.

For this practical reason and for the sake of simplicity, the discussion in this book is limited to large samples. As a rule, the standard error formulas given may become rather seriously inaccurate if applied to samples with fewer than 20 to 25 items, and are safest when used with much larger samples.[1]

5. Standard Error Formulas. *a. The Standard Error of a Frequency.*—As just shown, for a simple sample, the formula for the standard error of a frequency is

$$\epsilon_f = \sqrt{npq} = \sqrt{f\left(1 - \frac{f}{n}\right)}, \tag{113}$$

where p is the constant probability of drawing a success at any single draw, $q = 1 - p$, and f is the frequency in question. p refers to the probability in the universe, *i.e.*, to the true or expected probability, and f to the true frequency, but they are usually estimated from the sample when the latter is large.

We shall illustrate the use of this formula by application to the age distribution of an approximately simple sample of unemployed in New York City in 1930, shown in Table 67.

TABLE 67.—THE DISTRIBUTION OF UNEMPLOYED PERSONS BY AGE, IN A SIMPLE SAMPLE OF 100 UNEMPLOYED IN NEW YORK CITY, 1930

Age, years	Unemployed (f)	d	fd	fd^2
15–24...........................	28	−2	−56	112
25–34...........................	23	−1	−23	23
35–44...........................	21	0	0	0
45–54...........................	16	1	16	16
55–64...........................	9	2	18	36
65 and over....................	3	3	9	27
Total........................	100		−36	214

What is the range of error in the sample estimate of the relative number of unemployed in the age class 15–24? If we assume that the universe of the unemployed in New York is existent and large enough to be regarded as infinite for our purposes, and that it does not change appreciably during the process of sampling, then the probability of drawing an unemployed person in the age class 15–24 should be constant from draw to draw,

[1] See Sec. 6.

and from one sample of 100 persons to another. Thus the requirements of simple sampling are met, and we may determine the error of sampling of the frequency by formula (113). Since n is as large as 100, we accept 28 as an estimate of the true frequency in the age class 15–24. Substituting in formula (113),

$$\epsilon_{28} = \sqrt{28(1 - \tfrac{28}{100})},$$
$$\epsilon_{28} = 4.5.$$

In a large number of such samples the frequency in the age class 15–24 is approximately normally distributed, and the standard error just found is an estimate of the standard deviation of that normal distribution. Under these conditions, about two times in three the true frequency in the age class 15–24 will be
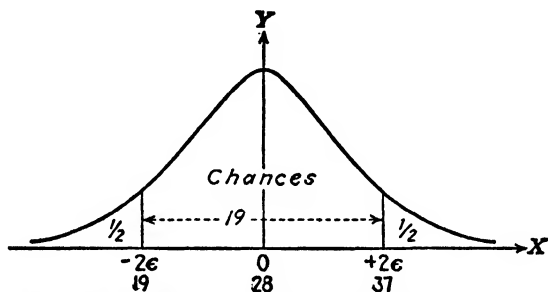
FIG. 52.—Showing a range within which the true number of unemployed in the age class 15–24 years will be enclosed about 19 times out of 20 (in the long run), as determined from a simple sample of 100 unemployed in New York City, 1930.

included within one standard error above and below the sample frequency of 28. That is, about two times out of three, we should expect the true number of unemployed persons in the age class 15–24 to be contained between the limits 28 ± 4.5, or between 23.5 and 32.5. If we want more security than this, we may multiply the standard error by two, getting limits of $28 \pm 2(4.5)$, or 19 to 37, within which about 19 times in 20 the true frequency will be found.[1] To attain practical certainty, we may multiply the error by three, giving chances of about 369 to 1 that the true frequency is enclosed between $28 \pm 3(4.5)$, or between 14.5 and 41.5. Usually, a range of twice the standard

[1] See Appendix Table 1.

error is regarded as safe enough. In case this range, here 19 to 37, seems too wide to be of much value, and it is wanted to narrow it, the size of the sample must necessarily be increased, since the size of the sampling error varies directly as $\sqrt{n}$ (see Sec. 6).

Evidently, if the size of the sample is decreased, the relative range of the sampling error increases, so that one reason why a small sample is not suitable for estimating the value of a parameter is easily seen. For example, suppose that the number of persons in the sample of Table 67 is only 10, and the frequency in the age class 15–24 is 3. Substituting in formula (113), with a factor $n/(n-1) = \frac{10}{9}$ inserted as a correction for the small size of the sample,

$$\epsilon_3 = \sqrt{\tfrac{10}{9}[3(1 - \tfrac{3}{10})]},$$
$$\epsilon_3 = 1.53.$$

We no longer have confidence in the sample frequency as an estimate of the universe frequency for use in the formula, but disregarding this, we find the range of twice the standard error to be $3 \pm 2(1.53) = -0.06$ to 6.06, or approximately 0 to 6. The ratio of twice the error to the frequency is now $3.06/3 = 1.02$, as compared with $\frac{9}{28} = 0.32$ for the larger sample.

If it is known that a sample was taken under Poisson conditions from a stratified universe, the standard error of a frequency estimated from it may be obtained by the formula

$$\epsilon_f{}^2 = n\bar{p}\bar{q} - n\sigma_{p_j}{}^2, \tag{114}$$

where p_j is the proportion of successes in any stratum, j, of the universe; $\bar{p}$ is the mean of the p_j's; $\sigma_{p_j}{}^2$ is the variance of the p_j's:

$$\sigma_{p_j}^2 = \sum^{k} \frac{f_j p_j^2}{n} - \bar{p}^2, \tag{115}$$

and k is the number of the strata. As in the case of formula (113), if the universe values of these statistics are not known, they are commonly estimated from the sample, provided the frequency in each stratum of the sample is fairly large (say, 50 or more).

TABLE 68.—ONE HUNDRED UNEMPLOYED PERSONS IN NEW YORK CITY, 1930, CLASSIFIED BY COLOR AND NATIVITY

| Age, years | Number of unemployed | | | |
	Total	Native white (1)	Foreign-born white (2)	Negro (3)
15–24........	28	5	18	5
25–34........	23	8	13	3
35–44........	21	9	7	5
45–54........	16	9	1	3
55–64........	9	8	0	3
65 and over..	3	3	0	0
Total......	100	42	39	19

Assume that Table 68 above is a Poisson sample, drawn as previously described, the strata being native white $(j = 1)$, foreign-born white $(j = 2)$, and Negro $(j = 3)$, as shown in the table. Let $n_1 = 42$, $n_2 = 39$, and $n_3 = 19$; and let the numbers in each stratum falling in the age class 15–24 be $f_1 = 5$, $f_2 = 18$, and $f_3 = 5$, giving $p_1 = \frac{5}{42} = 0.12$, $p_2 = \frac{18}{39} = 0.46$, and

$$p_3 = \frac{5}{19} = 0.26.$$

Then $\bar{p} = \frac{28}{100} = 0.28$, $\bar{q} = 1 - \bar{p} = 1 - 0.28 = 0.72$, and from formula (115)

$$\sigma_{p_j}{}^2 = [42(.12)^2 + 39(.46)^2 + 19(.26)^2]/100 - (.28)^2 = 0.023.$$

Substituting in formula (114),

$$\epsilon_f{}^2 = 100(.28)(.72) - 100(.023),$$
$$\epsilon_f{}^2 = 17.86,$$
$$\epsilon_f = 4.23.$$

Notice that this error is slightly smaller than that found on the assumption that Table 67 represented a simple sample.

It may be objected that in Table 68 the frequencies in cols. (1), (2), and (3) are not large enough to yield very good estimates of the true values in the universe.

b. The Standard Error of a Proportion.—In dealing with Table 67, above, as a simple sample we may think of the frequency 28 in the age class 15–24 as a proportion of the total frequency 100,

$p = \frac{28}{100} = 0.28$, and use formula (116) to find the standard error of this proportion:

$$\epsilon_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1 - p)}{n}}. \tag{116}[1]$$

Substituting in (116),

$$\epsilon_{.28} = \sqrt{\frac{.28(1 - .28)}{100}},$$
$$\epsilon_{.28} = 0.045.$$

Therefore the proportion of unemployed persons in the age class 15–24 estimated from the sample and the range of error of the true proportion may be written 0.28 ± 0.045. This means that the chances are two to one that the true proportion, or parameter, is not less than 0.235 or more than 0.325.

If we suppose, as we did in the preceding section, that Table 68 is a Poisson sample, the formula for the standard error of a proportion is

$$\epsilon_p{}^2 = \frac{\bar{p}\bar{q}}{n} - \frac{\sigma_{p_i}{}^2}{n}. \tag{117}$$

Using the same values as for formula (114), we have,

$$\epsilon_{.28}{}^2 = \frac{(.28)(.72)}{100} - \frac{.023}{100},$$
$$\epsilon_{.28}{}^2 = 0.001786,$$
$$\epsilon_{.28} = 0.0423,$$

which is again smaller than the standard error of the same proportion estimated from Table 67 regarded as a simple sample.

c. The Standard Error of an Arithmetic Mean.—Even when the universe departs considerably from normality, the means of large samples tend themselves to be normally distributed.

Formula (118) gives the standard error of the arithmetic mean found from a simple sample:

$$\epsilon_M = \frac{\sigma}{\sqrt{N}}, \tag{118}$$

[1] Just as a frequency is changed to a proportional frequency by dividing it by n, so the standard error of the former [formula (113)] is changed into the standard error of the latter [formula (116)] in the same way:

$$\epsilon_p = \frac{1}{n} \sqrt{npq} = \sqrt{\frac{pq}{n}}.$$

where N is the total frequency of the table or the sample, and σ is the standard deviation of the universe, estimated from the sample.

The mean of the simple sample of Table 67, taking the midpoint of the open interval at 70, is

$$M = A + i \cdot \frac{\Sigma fd}{N},$$
$$M = 40 + 10(-\tfrac{36}{100}) = 36.4.$$

The standard deviation is

$$\sigma = i \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2},$$
$$\sigma = 10 \sqrt{\frac{214}{100} - \left(\frac{-36}{100}\right)^2},$$
$$\sigma = 14.2.$$

Substituting these values in formula (118),

$$\epsilon_M = \frac{14.2}{\sqrt{100}},$$
$$\epsilon_M = 1.42.$$

We therefore write for the mean and its standard error

$$36.4 \pm 1.42.$$

For a Poisson sample, the standard error of the mean is given by the formula

$$\epsilon_M{}^2 = \frac{\sigma^2}{N} - \frac{\sigma_{m_j}{}^2}{N}, \tag{119}$$

where σ^2 is the variance of the universe estimated from the total sample, and

$$\sigma_{m_j}{}^2 = \frac{\sum_{}^{k} f_j m_j{}^2}{N} - M^2, \tag{120}$$

where m_j is the mean of the jth stratum. As usual, all statistics are estimated from the sample when the true values in the universe are not known.

Referring to Table 68, we found that

$$\sigma^2 = (14.2)^2 = 201.64,$$

and $M^2 = (36.4)^2 = 1,324.96$. Let the mean age of the native whites be $m_1 = 44$, of the foreign-born whites $m_2 = 27.7$, and of the Negroes $m_3 = 38$. We compute

$$\sigma_{m_i}^2 = \frac{42(44)^2 + 39(27.7)^2 + 19(38)^2}{100} - 1,324.96,$$

$$\sigma_{m_i}^2 = 61.76.$$

Substituting in formula (119),

$$\epsilon_M^2 = \frac{201.64}{100} - \frac{61.76}{100} = 1.40,$$

$$\epsilon_M = 1.18.$$

As before, we see that the standard error of the Poisson sample is smaller than that of the simple sample.

The standard error of the mean is most useful in testing the significance of the difference between two means, to be treated later.

d. Standard Error of the Standard Deviation.—For a simple sample drawn from an approximately normally distributed universe, the standard error of a standard deviation is

$$\epsilon_\sigma = \frac{\sigma}{\sqrt{2N}} \tag{121}$$

where σ is the standard deviation of the universe, estimated from the sample.

TABLE 69.—SCORES OF 100 COMMUNITIES ON A COMMUNITY ORGANIZATION TEST

Score (X)	Communities (f)	d	fd	fd²	Accumulated frequency	X − M	f(X − M)
80–99	9	2	18	36	100	41.4	372.6
60–79	17	1	17	17	91	21.4	363.8
40–59	43	0	0	0	74	1.4	60.2
20–39	20	−1	−20	20	31	18.6	372.0
0–19	11	−2	−22	44	11	38.6	424.6
Total...	100		− 7	117			1,593.2

Table 67 is a J-shaped rather than a normal distribution, so it does not lend itself to formula (121). We shall, however, risk applying the formula to Table 69, which is only moderately

skewed.   The 100 communities were taken at random from the total of some 300 cities of a given size class in the United States, the name of each community taken being replaced before the next draw.   The sample may, therefore, be regarded as a simple sample, representing an infinite existent universe of cities like the 300 cities reported by the census.

The standard deviation of Table 69 is

$$\sigma = 20 \sqrt{\frac{117}{100} - \left(\frac{-7}{100}\right)^2} = 21.6.$$

So that, by formula (121),

$$\epsilon_\sigma = \frac{21.6}{\sqrt{(2)(100)}} = 1.53.$$

And we write for the standard deviation and its standard error 21.6 ± 1.53.

e. Standard Error of a Variance.—Assuming as before a simple sample from an approximately normal universe, the variance, σ^2, has the standard error,

$$\epsilon_{\sigma^2} = \sigma^2 \sqrt{\frac{2}{N}}. \tag{122}$$

The variance of Table 69 is $(21.6)^2 = 466.56$, and its standard error is

$$\epsilon_{\sigma^2} = 466.56 \sqrt{\tfrac{2}{100}} = 65.78.$$

f. Standard Errors of Sampling from a Limited Universe.—A great part of the sampling done in social research is from limited rather than from infinite universes.   It has already been seen that from a limited universe a random sample can be drawn, but a simple sample cannot.   All the formulas for finding the standard errors of a simple sample given above, therefore, need a correction if the sample is drawn at random from a limited universe, *i.e.*, if the sample is random but not simple.   In the case of a mean, frequency, or proportion, the correction consists in applying the multiplying factor, $\sqrt{(U - s)/U}$, to the standard error of a simple sample, where U is the number of events in the limited universe, and s is the number of events in the sample, so that $s = n$ or N in our formulas above.   It is not certain that this correction is applicable to standard errors other than those mentioned.

One or two illustrations will suffice to show how the correction factor $\sqrt{(U - s)/U}$ is used. In the section dealing with the standard error of a frequency for Table 67, we found $\epsilon_{28} = 4.5$. Now if we regard the universe of the unemployed in New York City from which this sample of 100 was drawn as a limited universe, consisting in the year 1930 of an average of 300,000 persons, we have

$$\sqrt{\frac{(U - s)}{U}} = \sqrt{\frac{(300,000 - 100)}{300,000}} = 0.9997.$$

Multiplying this into the standard error found by assuming an infinite universe we get $.9997(4.5) = 4.499$, which for all practical purposes is the same as before. This suggests that when the limited universe is quite large there is no need to make the correction.

Suppose from a limited universe of 1,382 divorces granted in a certain court in 1939, a random sample of 200 is drawn. From this sample the mean legal cost of getting a divorce is found to be \$136, and the standard deviation \$32. If we regard the universe as limited, the standard error of \$136 is obtained by multiplying the corrective factor into formula (118), the standard error of the mean of an infinite sample:

$$\sqrt{\frac{(U - s)}{U}} \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{1,382 - 200}{1,382}} \left(\frac{32}{\sqrt{200}}\right) = .925(2.26) = 2.09.$$

In this case, the correction for a limited universe reduces the standard error of the mean over 7.0 per cent.

g. Standard Errors When the Unit of Sampling Is a Group of Events, or District; and the Standard Error of a Population Rate.— When a sample of districts, instead of individual events, is taken, the district simply replaces the individual event in the appropriate standard error formula. That is, n or N becomes the number of districts, rather than the number of events. The proper standard error formula to use in any given case depends as before on the conditions under which the sample was drawn. However, only those standard error formulas are appropriate for districts that apply to variables, because, disregarding sampling errors within districts, each district is merely one value of a variable, such as a proportion or mean, determined by the events within the district. In finding the mean, the

variance, and so on, of the district values, it is usually advisable to weight the latter by the number of events in the respective districts.

TABLE 70.—BIRTH RATES IN 20 COUNTIES OF WISCONSIN, 1935

County	Birth rate = (p_i)	Population, 1930 = (Y_i)	Products = $(Y_i p_i)$	Squares = $(Y_i p_i^2)$
1	.0186	8,003	148.856	2.768,718
2	.0222	21,054	467.399	10.376,253
3	.0184	34,301	631.138	11.612,947
4	.0125	15,006	187.575	2.344,688
5	.0221	70,249	1,552.503	34.310,314
6	.0175	15,330	268.275	4.694,813
7	.0172	10,233	176.008	3.027,331
8	.0157	16,848	264.514	4.152,864
9	.0205	37,342	765.511	15.692,976
10	.0173	34,165	591.055	10.225,243
11	.0174	30,503	530.752	9.235,088
12	.0225	16,781	377.573	8.495,381
13	.0171	112,737	1,927.803	32.965,426
14	.0144	52,092	750.125	10.801,797
15	.0208	18,182	378.186	7.866,260
16	.0162	46,583	754.645	12.225,243
17	.0187	27,037	505.592	9.454,569
18	.0220	81,087	903.914	19.886,108
19	.0178	3,768	67.070	1.193,853
20 = n	.0173	59,883	1,035.976	17.922,383
Total.......		671,184	12,284.470	229.252,255

In Table 70 is a random sample of 20 counties of Wisconsin, showing their birth rates ($p_i = X_i/Y_i$, where X_i = births) in 1935. The mean birth rate for the table is

$$\bar{p} = \frac{\sum\limits_{}^{n} (Y_i p_i)}{\sum\limits_{}^{n} Y_i} = \frac{12{,}284.470}{671{,}184} = .01830, \qquad (123)$$

and the variance is

$$\sigma_p{}^2 = \frac{\sum\limits_{}^{n} Y_i p_i{}^2}{\sum\limits_{}^{n} Y_i} - \left(\frac{\sum\limits_{}^{n} Y_i p_i}{\sum\limits_{}^{n} Y_i}\right)^2 = \frac{229.252{,}255}{671{,}184} - \left(\frac{12{,}284.470}{671{,}184}\right)^2$$

$$= .00000667, \quad \text{so that} \quad \sigma_p = \sqrt{.00000667} = .0025826.$$

$$(124)$$

By formula (118),

$$\epsilon_M = \frac{.0025826}{\sqrt{20}} = .0005775.$$

Or, combining formulas (124) and (118), and adding a term due to errors of sampling within a district, the standard error of a population rate is approximately

$$\epsilon_{\bar{p}}^2 = \frac{1}{n}\left[\frac{\sum\limits_{}^{n} Y_i p_i^2}{\sum\limits_{}^{n} Y_i} - \left(\frac{\sum\limits_{}^{n} Y_i p_i}{\sum\limits_{}^{n} Y_i}\right)^2\right] + \frac{\bar{p}\bar{q}}{\sum\limits_{}^{n} Y_i}. \qquad (125)[1]$$

Thus, for Table 70,

$$\epsilon_{\bar{p}}^2 = (.0005775)^2 + \frac{.0183(.9817)}{671,184}$$

$$= .0000003335 + \frac{.017,965,110}{671,184} = .0000003335$$

$$+ .000000026766 = .000000360266,$$

$$\epsilon_{\bar{p}} = 0.0006.$$

Since we think of the 71 counties of Wisconsin in 1935 as a limited universe of birth rates, we should apply the correction[2] factor, $\sqrt{(71 - 20)/71} = \sqrt{.7183}$, giving

$$.0006(.8475) = 0.000509$$

as the final standard error. We therefore write the mean birth rate and its standard error: 0.0183 ± 0.00051, or multiplied by

[1] More exactly, the last term is

$$\frac{1}{\sum\limits_{}^{n} Y_i}\left(\bar{p} - \frac{\sum\limits_{}^{n} Y_i p_i^2}{\sum\limits_{}^{n} Y_i}\right)$$

$$= \frac{1}{671,184}\left(.0183 - \frac{229.252,255}{671,184}\right) = 0.000000026756.$$

When the population is large, however, this term is usually negligible.

[2] Or, using population weights, the correction factor is

$$\sqrt{\frac{\sum\limits_{}^{N} Y_i - \sum\limits_{}^{n} Y_i}{\sum\limits_{}^{N} Y_i}},$$

where N is the number of districts in the universe and n is the number of districts in the sample.

1,000, 18.30 ± 0.51. The chances are 19 in 20 that the birth rate per 1,000 for the state as a whole will be enclosed within the sample range 18.30 ± 2(.51) = 17.28 to 19.32. As a matter of fact, the birth rate for Wisconsin in 1935 was 17.3, almost at the bottom limit. This is because the city of Milwaukee, with a very low birth rate of 15.0, happened to be left out of the sample. The birth rate for the state without Milwaukee was 18.09, which is well within the estimated range of sampling error.

This case illustrates one of the dangers of sampling by districts, namely, that the sample may omit a district with an extreme value and a very large number of events. This is avoided when the events are sampled directly. In the case of birth rates and other population rates, sampling by districts is unavoidable, but counties like Milwaukee should be subdivided into several average-sized population districts, each with the given Milwaukee birth rate. Then the chance of such an omission from the sample is lessened.

It was assumed above that all the events in each sample district were used to determine the district value. Sometimes it is necessary to sample the events in sample districts. This might be the case if we wanted to study a few hundred farmers' household accounts in a given state. We would probably draw a sample of counties, but could not get accounts from all of the farmers in a sample county. This random sampling of events would increase the sampling error within the districts. For example, in formulas (123), (124), and (125), Y_i would become y_i, where y_i is the size of the sample population drawn from district i, on which the birth rate, p_i is calculated.

6. Control of Sampling Error by Size of Sample.—The number of items that a sample should contain to yield a satisfactorily accurate picture of the distribution in the universe from which it is taken depends on the number of different kinds or classes of items that it is necessary to distinguish (*i.e.*, on the heterogeneity) in the universe, on the relative frequencies of the items in each class, and on whether the items are *stratified* or mixed. This may be explained by a simple illustration.

If the universe is limited and consists of two individuals, a white and a Negro, who are to be examined for skin color, evidently the sample will fall short of giving a proper picture of the universe, or of being *representative*, unless it contains

both of the individuals, or *all* of the items in the universe. Should the universe contain thousands of individuals but only two skin colors, each equally distributed among the population and subject to no variation in shade from one individual to another, a perfectly representative sample need still contain only one individual of a color, each drawn at random from a color group or stratum.   If the same universe is not stratified, however, but the sample has to be drawn at random from the two races mixed together, a sample of more than two individuals should be taken, since otherwise the chance of getting all individuals of the same color is one in four $[(\frac{1}{2} + \frac{1}{2})^2 = (\frac{1}{4} + \frac{1}{4}) + \frac{1}{2}]$. In fact, a fairly large sample—say not less than 25 items—is now advisable, to lessen the risk that one of the colors will appear much more often than the other, and so give a false impression of its relative frequency in the universe.   Finally, suppose that the universe includes many individuals of the same race—say Negro—but the skin color varies widely among the individuals.   Suppose, further, that we want to learn from a sample the relative frequency of occurrence of the various shades of skin color, including the extreme shades that the color takes. If some shade, say intensely black, exists in only one individual per 1,000 in the universe, a random sample containing even as many as 100 individuals will fail to include it nine times in 10 $[(1.000 - 0.001)^{100}]$.

If it is wanted to use the sample merely to estimate the mean of the universe distribution, omissions at one part of the scale may cancel omissions at another part, so that the size of the sample need not be so great.   Yet, for a given degree of accuracy in the estimate, the size of the sample must be increased as the variance, σ^2, of the universe distribution, also estimated from the sample, increases.

It is theoretically a simple matter to reduce the standard error to any desired value by merely increasing the size of the sample, N.   For this purpose, we have the formula

$$N_2 = a^2 N_1, \tag{126}$$

where

$$a = \frac{\epsilon_1}{\epsilon_2}. \tag{127}$$

ϵ_1 is the value of the original standard error, ϵ_2 is the value of the

desired standard error, N_1 is the size of the original sample, and N_2 is the size of sample needed to reduce ϵ_1 to ϵ_2.

In Sec. 5c, above, we found the mean, 36.4, and its standard error, 1.42, from a simple sample of 100 items. What size of sample is required to reduce the standard error to one-half its present value? According to this requirement, $\epsilon_2 = \epsilon_1/2$, so that $a = 2$. Substituting in formula (126),

$$N_2 = (2)^2(100) = 400.$$

Notice that when the divisor of the original standard error is named, we have only to multiply the size of the sample by the *square* of that divisor. That is, to divide the standard error by two, we multiply the size of the sample by four. This rule applies to any common standard error except the standard error of a frequency. The easiest way to deal with the standard error of a frequency is to substitute for it the standard error of the equivalent proportion, for which the above rule holds. For example, in Sec. 5b, above, the frequency, 28, in Table 67 was changed to the proportion, 0.28, for which the standard error was estimated to be 0.045. To reduce this error, or the relative error of the corresponding frequency, to one-third of its value, we multiply $N_1 = 100$ by $(3)^2 = 9$, giving 900 as the size of the sample required.

The problem of determining the proper size of fairly large samples may be approached in terms of *confidence* or *fiducial limits*. That is, we may require the sample to be of such a size that about $2P$ times in 100 the value of a parameter in which we are interested will be enclosed within a specified range. Using again the example of Sec. 5c, let us say that it is wanted to take a sample of such size that the chances are about 95 in 100 that the parameter will be enclosed within a range that extends on either side of S a distance equal to 10 per cent of the value of S. We then require

$$\epsilon_s\left(\frac{x}{\sigma}\right) = p'S, \tag{128}$$

where, for this particular problem, $S = M_x = 36.4$,

$$\epsilon_{Mx} = \sigma/\sqrt{N_2} = 14.2/\sqrt{N_2},$$

x is a mean deviate of S, p' is one-half the width of the range expressed as a percentage of the value of S, $\frac{x}{\sigma}$ is the value read from a table of normal areas corresponding to one-half of the area

enclosed by the specified confidence limits, $P = .95/2 = 0.475$, and N_2 is the required size of the sample. Assuming that the distribution of sample means from large samples is approximately normal in form, we turn to Appendix Table 1, and find that $2P = 2(.475) = 0.95$ between the points $x/\sigma = \pm 1.96$, so that $x/\sigma = 1.96$. Substituting in formula (128), $\dfrac{\sigma}{\sqrt{N_2}} \cdot \dfrac{x}{\sigma} = p'\bar{X}$, we have

$$\frac{14.2}{\sqrt{N_2}} (1.96) = .10(36.4),$$

$$\sqrt{N_2} = 7.64,$$
$$N_2 = 58.4.$$

To check this, we write

$$S \pm \epsilon_s \left(\frac{x}{\sigma}\right), \qquad (129)$$

$$36.4 \pm \frac{14.2}{\sqrt{58.4}} (1.96),$$

$$36.4 \pm 3.64.$$

Since 3.64 is 10 per cent of 36.4, we have the result desired.

Notice, as a further check, that in our solution the standard error of the mean is only $14.2/\sqrt{58.4} = 1.86$. If the mean age
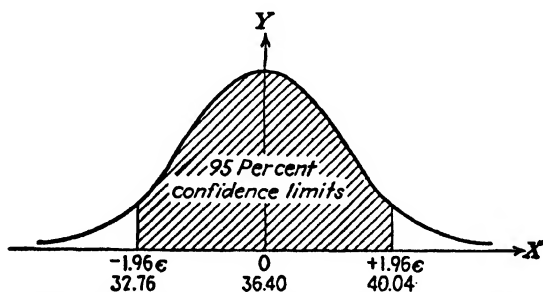
Fig. 53.—Showing 95 per cent confidence limits for the mean of a random sample of 58 items. About ninety-five chances in a hundred the true mean will be enclosed within the limits 32.76 and 40.04.

varies by 10 per cent of its value, however, it will vary by ± 3.64 years, which is $3.64/1.86 = 1.96\epsilon_M$. But ordinates of the normal curve at the points ± 1.96 standard errors include 95 per cent of the area of the curve. Therefore, the chances are about 95 in 100 that the true mean will be found within ± 10 per cent of the value of the mean of our sample.

7. Error in Mean vs. Individual Predictions from a Regression Equation.—Interest often centers in predicting *averages* rather than individual values from a regression equation.[1]   Thus, out of 20 counties with birth rates of 18 per 1,000, how far does the *mean observed* death rate differ from the *most probable* death rate predicted by the regression equation?   The scatter of the observed means of such samples around the predicted value in this case depends upon the size of the sample, N, as well as upon the value of r, and may be found from the equation

$$\epsilon_{\bar{p}} = \frac{S_y}{\sqrt{N}}.$$ (130)

It appears from this that the scatter in predicting the mean value of Y corresponding to a given value of X, or to the mid-point of an X class interval, is reduced, compared to the scatter in calculating any individual value of Y, in proportion to the square root of the size of the sample from which the mean is found.   For the data of Table 50 of Chap. X, if we take a sample of 20 counties all with approximately the same birth rate, equation (130) gives

$$\epsilon_{\bar{p}} = \frac{1.86}{\sqrt{20}} = 0.416,$$

which is less than a quarter of the size of the standard error of estimate, S_y ($= 1.86$), that governs the prediction of a death rate from a birth rate in the case of a single county.

8. Representativeness of a Sample.—The test of the goodness of a sample is simply the test of its representativeness.   If we knew the value of the parameter, we could measure the representativeness of the sample in terms of the percentage deviation of the statistic from the parameter.   Thus, if s is the statistic and S is the corresponding parameter, the formula for measuring the representativeness, Rp, is

$$Rp = \left[100 - 100 \frac{(s - S)}{S} \right] \text{ per cent,}$$ (131)

where we take $S - s$ if $S > s$.

The value of the parameter is seldom known, however, for if it were, it is not likely that a sample would be taken.   This

[1] See Chap. X, Tables 50 and 51.

means that there is generally no direct way of measuring the representativeness of a sample. The nearest approximation would be to take several additional samples, each equivalent in method of drawing and (preferably) in size to the original sample. Then, in addition to noting the variation of certain statistics from one of these samples to another, we might pool the samples to obtain an average statistic, average the statistics found from them, and substitute this average value for S in formula (131), above. But if this were done, we would of course at once abandon the statistic from the original sample in favor of the average statistic from the several samples, whose representativeness would still be unknown. As a rule, therefore, the best that we can do in the way of formulating an index of representativeness is to rely on a large sample, and, where possible, stratification of the universe, and measure the probable maximum deviation of the statistic from the parameter in terms of, say, two standard errors of the statistic (ϵ_s). This permits us to say that the *probable minimum* representativeness, $\widetilde{Rp}$, of the statistic is

$$\widetilde{Rp} = \left(100 - \frac{200\epsilon_s}{s}\right) \text{ per cent.} \tag{132}$$

In Sec. 5c, above, we found the mean age of a simple sample of 100 ages to be 36.4 years, and the standard error of this mean to be 1.42 years. If we knew that the mean age in the universe from which the sample was taken was 37.5 years, we would find the representativeness of the sample by formula (131) to be

$$\begin{aligned}
R\dot{p} &= \left[100 - 100\,\frac{(37.5 - 36.4)}{37.5}\right] \\
&= 100 - 2.9 \\
&= 97.1 \text{ per cent.}
\end{aligned}$$

But if we did not know the parameter value, we would estimate the probable minimum representativeness by formula (132),

$$\begin{aligned}
\widetilde{Rp} &= \left[100 - 200\left(\frac{1.42}{36.40}\right)\right] \\
&= 100 - 7.8 \\
&= 92.2 \text{ per cent.}
\end{aligned}$$

An indirect but important method of judging the representativeness of a sample makes use of the circumstance that although

the value of the attribute or variable for which we are sampling is not known in the universe, other universe values may be known; and if the sample can be shown to be representative of the latter, it is likely to be representative of the former also.   As an illustration of this, we may draw a random sample of families in an Alabama county for the purpose of determining by field visits the percentage whose annual income falls below a certain minimum level.   After the sample is obtained, it may be compared with the figures of the latest Federal census for the given county in regard to median size of family, the proportions of families having different numbers of children under 10 years of age, the percentage of families that do not own their home, and the median rental paid.   If a reasonably close agreement is found between the sample and the census population in these respects, the sample may usually be regarded as satisfactory also for the study of incomes.

Exercises

1. Define in both time and space (1) an infinite universe, (2) a limited universe, (3) a hypothetical universe, choosing in each case a universe of interest to social scientists.

2. Give an example of a universe of social attributes, and define the event and the "success."

3. Illustrate a universe of a social variable.

4. Draw a sample of events or values from an actual known social universe, so that the sample will be (1) random, (2) simple, (3) Poisson (stratified).

5. Draw a random sample of districts from a known social universe of your own choosing.

6. In Table 34 of Chap. VIII, what is the standard error of the frequency in the class $X = 0$?   What does it mean?

7. In Table 34 of Chap. VIII, what is the standard error of the proportion of prisoners with no previous arrests?   How does this standard error compare with that for a frequency found in Exercise 6 above?

8. Ten thousand marriage certificates issued in the same month in five large American cities are taken as the universe, and a random sample of 500 certificates is drawn from them.   After one year, it is found that 78 of the 500 marriages are divorced.   What is the mean probability of divorce in this heterogeneous universe of marriages, and what is its approximate standard error?

9. Judging from the sample in the following table, what is a range within which the true number of Orientals immigrating to the United

Year	Chinese	Japanese	Total
1929	102	65	167
1928	115	49	164
1927	105	43	148
1925 and 1926	187	74	261
Total...............	509	231	740

States in the year 1929, expressed as a percentage of the total Oriental immigration over the five-year period, 1925–1929, will fall 95 times out of 100? Compare the standard errors found on the assumptions of a simple sample from an infinite universe, a random sample from a limited universe (total Chinese immigrants, 5,090, total Japanese immigrants, 2,314), and a Poisson sample from a limited universe. If the sampling was random and proportional between Chinese and Japanese, which of these assumptions seems preferable, and why? ·

10. What is the standard error of the mean in the table of Exercise 3 of Chap. XIII for urban families? How do you interpret it?

11. Below are given the number of children under 5 years of age and the number of women aged 15–45 years for each of 20 random counties in Wisconsin in 1930, with the resulting fertility ratios.

Within what range will the fertility ratio for the state of Wisconsin fall, 95 times out of 100? (NOTE: the fertility ratio in the State of

FERTILITY RATIOS AND BASIC DATA, 20 RANDOM COUNTIES IN WISCONSIN, 1930*

County code	Children under 5 = X_i	Women 15–45 = Y_i	$\dfrac{X_i}{Y_i} = p_i$
1	731	1,523	.48
2	1,968	4,331	.45
3	3,463	7,084	.49
4	1,243	2,723	.46
5	6,998	16,408	.43
6	1,562	3,157	.49
7	953	1,924	.50
8	1,619	3,526	.46
9	3,707	8,118	.46
10	3,330	6,765	.49
11	2,536	6,166	.41
12 .	1,745	3,339	.52
13	10,016	27,401	.37
14	4,504	10,889	.41
15	1,757	3,671	.48

* From Fifteenth Census of the United States, 1930, Population, Vol. III, Part 2, pp. 1314–1319.

County code	Children under $5 = X_i$	Women 15–45 = Y_i	$\dfrac{X_i}{Y_i} = p_i$
16	3,707	10,546	.35
17	2,796	5,550	.50
18	3,758	9,692	.39
19	395	648	.61
20	5,364	13,330	.40
Total............	62,152	146,791	.42

* From Fifteenth Census of the United States, 1930, Population, Vol. III, Part 2, pp. 1314–1319.

Wisconsin as a whole in 1930 was about 0.41. There are 71 counties in the state.)

12. Within what range will the standard deviation of the fertility ratios in the universe fall 95 times in 100, according to the random sample in the table of Exercise 11 above? Can the standard error of the standard deviation be applied to urban families in the table of Exercise 3 of Chap. VIII? Explain.

13. What size sample of rural nonfarm families in the table of Exercise 3 of Chap. XIII is needed to reduce the standard error to one-half its value?

14. In the table of Exercise 9 above, what size sample of Japanese is required to confine the true value of the proportion of immigrants in the year 1929 within 5 per cent of the observed value 99 times in 100 (*i.e.*, within 99 per cent confidence limits)?

15. Measure the probable minimum representativeness of the mean score in Table 69, above.

References

BATEN, W. D.: *Mathematical Statistics*, Chaps. XI and XV, John Wiley & Sons, Inc., New York, 1938.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chaps. XII and XIII, Prentice-Hall, Inc., New York, 1939.

MILLS, F. C.: *Statistical Method*, rev. ed., Chaps. XIV and XVIII, Henry Holt and Company, Inc., New York, 1938.

PETERS, C. C., and W. R. VAN VOORHIS: *Statistical Procedures and Their Mathematical Bases*, Chaps. V, VI, and XIV, McGraw-Hill Book Company, Inc., New York, 1940.

TIPPETT, L. H. C.: *The Methods of Statistics*, 2d ed., Chaps. II, III, IV, and VIII, Williams and Norgate, Ltd., London, 1937.

TRELOAR, A. E.: *Elements of Statistical Reasoning*, Chaps. X, XI, XIV, and XV, John Wiley & Sons, Inc., New York, 1939.

YULE, G. U., and M. G. KENDALL: *An Introduction to the Theory of Statistics*, Chaps. XVIII–XXII, Charles Griffin & Company, Ltd., London, 1937.

CHAPTER XIII

THE SIGNIFICANCE OF DIFFERENCES

1. The Meaning of Tests of Significance.—It has been seen that the value of a statistic estimated from a random sample usually differs somewhat from the true value, or parameter, in the universe from which the sample is drawn. Similarly, the values of a statistic, such as the mean, yielded by two or more random samples from the same universe, will almost never be exactly the same, and may sometimes be quite far apart. Such variations, however, are due merely to chance errors of sampling and imply no actual differences. On the other hand, samples from different universes yield statistics of different values which represent real differences in the parameters. It therefore becomes a matter of great importance in investigations based on sampling to distinguish between real differences and accidental ones.

If we could be certain that two or more samples were taken at random from the same universe or from different universes, there would, of course, be no problem. In most of the practical sampling work done in the social sciences, however, the investigator cannot feel entirely confident that his samples are random, and he knows so little about the universes from which they are taken that he cannot say whether these universes are essentially the same or different. For example, if we try to take random samples of 500 persons each from the total population of a city like Chicago, it will not be easy to insure that the selection will be random, or even to guarantee that the persons drawn will all belong to the population of Chicago. If the several samples are not taken on the same day or even at the same hour, the populations sampled may be radically different, because of the traffic in and out of the city in the mornings and evenings, on week ends and holidays. As a consequence of such uncertainties, an investigator feels the need for some kind of test that will lend additional security to any inferences that he may draw from samples. The development of such tests, based on the mathe-

matical theory of probability, constitutes the major part of present-day statistical method.

In Chap. XII, it was usually assumed that if the sampling was random or simple from a normally distributed universe, the statistic itself would be normally distributed over many samples. By the use of the standard error, it then became possible to estimate from the normal curve the probability that the parameter would be enclosed within specified limits. It is now necessary only to extend these ideas to the *differences* between statistics, and to direct attention to the common rule that if a difference as large as the one observed might occur by chance no oftener than five times in 100, it is regarded as a real difference. In that case, the difference is said to be *significant.*

The differences that are tested by this method are of two general kinds. The first is the difference between the value of a statistic and the value of a known or hypothetical parameter. For example, can a group of mothers with a mean age of 27 years be a random sample from a universe of mothers whose mean age is 24 years? Or, can a correlation coefficient, $r = .34$, be a random statistic from a universe in which there is no correlation, *i.e.,* where $r = 0$? The second kind of difference that is frequently dealt with is the difference between the statistics from two or more samples. Thus, can two groups of mothers, one with a mean age of 27 years and the other with a mean age of 31 years, be random samples from the same universe? If the test shows that the difference is significant, it is inferred that the answer to the above questions is negative, on the grounds that a negative answer is highly probable. If the test fails to show a significant difference, the sample is regarded as a random sample from a given universe, or two samples are regarded as random samples from the same universe, until tests applied to larger samples show the contrary. If it is not positively known that the samples are random, a nonsignificant test at least allows us to say that the observed differences are no greater than might occur with random samples.

If a difference is defined as real when the probability of its occurring by chance is as low as five in 100, we are said to be using the 5 *per cent level of significance.* The fixing of this critical probability is arbitrary and a matter of convention. The 5 per cent level is rather widely used at present, but the 1 per cent

level is preferred when there is need to be more conservative. Reference to Appendix Table 1 will show that 5 per cent of the area of the normal curve lies beyond ordinates erected at about two standard errors on each side of the mean, while 1 per cent of the area falls beyond ordinates at plus and minus 2.58 standard errors (see Fig. 54). It was formerly the practice to insist that an observed difference must fall as far out as three standard errors. At that point the probability is only about 27 in 10,000 that so large a difference might occur by chance in either direction. This is too stringent for ordinary purposes, because it causes the investigator to withhold judgment in an unnecessarily large proportion of cases.
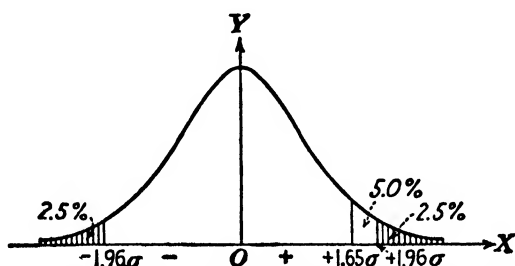
Fɪɢ. 54.—Five per cent of the area of the normal curve taken at the positive end only, and divided equally between the positive and negative ends.

Notice that since the 5 per cent level of significance, for example, includes 2.5 per cent of the area of the normal curve at each end of the X scale, it implies that the probability of getting *either* a positive or a negative difference is sought. If it is desired to find the probability of getting say a positive difference only, the reading is limited to the positive end of the scale (see Fig. 54).

2. The Significance of a Correlation Coefficient.—Suppose we have a correlation coefficient $r = .34$ from a simple sample of 30 pairs of variates from normal universes, one variate being scores on an I.Q. test and the other the scores of the same individuals on a personality test. Is the value of the observed r here so small that it might occur as a random error in a sample from a universe in which there is no correlation?

To answer this question, we test the difference of the observed value of r from zero. Appendix Table 4 has been designed to provide a ready-made test of this sort in the case of the correla-

tion coefficient. From it the value of the coefficient that is just significantly different from zero at the 5 per cent (or the 1 per cent) level of significance may be read off at once, and compared with the observed value. The table is entered with $N - 2$ degrees of freedom, which in this case are $30 - 2 = 28$. At the 5 per cent level we find that an $r = .36$ is just significant. Since the value of our r ($\pm.34$) is slightly smaller than this, it might occur by chance a little oftener than five times in 100. If we are governed strictly by the 5 per cent criterion, therefore, we cannot accept an $r = .34$ as significantly different from zero.

For simple samples from a normal universe so large that N is not covered by Appendix Table 4, formula (133) is convenient to test the hypothesis that the observed value of r is not different from zero.

$$\epsilon_r = \frac{1}{\sqrt{N}}. \tag{133}$$

In the example above,

$$\epsilon_r = \frac{1}{\sqrt{30}} = 0.18.$$

The ratio of the statistic r to its standard error, called the *critical ratio* (*C.R.*), is

$$C.R. = \frac{.34}{.18} = 1.89.$$

Entering a table of normal areas (Appendix Table 1) with $C.R. = x/\sigma = 1.89$, the probability is found to be about six in 100 that a larger value of r than that observed might occur because of random errors of sampling. Again we find that the value $r = .34$ is not quite significantly different from zero. It might have come from a universe in which there was no correlation at all.

3. The Significance of g_1 and g_2.—In a problem in Chap. IX we found the measure of skewness of a certain distribution to be $g_1 = 1.17$. The standard error of g_1 in large samples is approximately

$$\epsilon_{g_1} = \sqrt{\frac{6}{N}}. \tag{134}$$

Substituting in formula (134),

$$\epsilon_{g_1} = \sqrt{\tfrac{6}{252}} = 0.154.$$

If now we divide the value of g_1 by its standard error, we get the critical ratio $1.17/0.154 = 7.6$. Since this is much more than two standard errors, chance is ruled out, and we conclude that the distribution could not reasonably be regarded as a random sample from a normal universe.

Let us next test the value of the measure of kurtosis, $g_2 = 0.86$, found for the same distribution in Chap. IX. The standard error of g_2 for large samples is

$$\epsilon_{g_2} = \sqrt{\frac{24}{N}}. \tag{135}$$

For $N = 252$,

$$\epsilon_{g_2} = \sqrt{\tfrac{24}{252}} = 0.309.$$

The critical ratio is therefore $0.86/0.309 = 2.78$. Thus the peaked distribution in question could not have been drawn at random from a normally distributed universe.[1]

4. The Significance of the Difference between Any Two Statistics.—The variance of the differences, D, between n paired values of two variables, X_1 and X_2, is, by the usual formula,

$$\sigma_D{}^2 = \frac{\Sigma(D - M_D)^2}{N},$$

where M_D is the mean of the differences. Or, since

$$D = X_1 - X_2, \quad \text{and} \quad M_D = \frac{\Sigma D}{N} = \frac{\Sigma(X_1 - X_2)}{N},$$

$$\sigma_D{}^2 = \frac{\sum \left[(X_1 - X_2) - \dfrac{\Sigma(X_1 - X_2)}{N} \right]^2}{N}$$

$$= \frac{\sum \left(X_1 - X_2 - \dfrac{\Sigma X_1}{N} + \dfrac{\Sigma X_2}{N} \right)^2}{N}$$

$$= \frac{\Sigma[(X_1 - M_{X1}) - (X_2 - M_{X2})]^2}{N}.$$

---

[1] For a more exact interpretation of the critical ratio g_2/ϵ_{g_2}, see L. H. C. Tippett, *The Methods of Statistics*, 2nd. ed., page 86.

Letting $X_1 - M_{X_1} = x_1$, and $X_2 - M_{X_2} = x_2$,

$$\sigma_D{}^2 = \frac{\Sigma(x_1 - x_2)^2}{N}$$

$$= \frac{\Sigma(x_1{}^2 - 2x_1x_2 + x_2{}^2)}{N}$$

$$= \frac{\Sigma x_1{}^2}{N} - \frac{2\Sigma x_1x_2}{N} + \frac{\Sigma x_2{}^2}{N}$$

$$= \sigma_1{}^2 + \sigma_2{}^2 - 2\sigma_1\sigma_2 \frac{\Sigma x_1x_2}{N\sigma_1\sigma_2}.$$

By formula (81), $\Sigma x_1x_2/N\sigma_1\sigma_2 = r_{12}$, so that

$$\sigma_D{}^2 = \sigma_1{}^2 + \sigma_1{}^2 - 2r_{12}\sigma_1\sigma_2.$$

If now we let $\sigma = \epsilon$, we have

$$\epsilon_D{}^2 = \epsilon_1{}^2 + \epsilon_2{}^2 - 2r_{12}\epsilon_1\epsilon_2, \tag{136}$$

where ϵ_1 is the standard error of the statistic in the first sample, ϵ_2 is the standard error of the corresponding statistic in the second sample, and r_{12} is the correlation coefficient between a number of sample values of the two statistics.

Usually, correlation between two sample statistics is purposely introduced by drawing one sample at random, and then matching on some principle each of the items or values so drawn with an item or value from another population. For example, the I.Q.'s of a random sample of criminals may be matched or paired with the I.Q.'s of their brothers.

If the statistics are the means or proportions from two samples whose individual values are matched in some way, the simple correlation coefficient, r_{12}, in the case of means, or r_t (tetrachoric correlation coefficient) in the case of proportions, may be used to determine the amount of correlation between the paired items of the two samples. Where correlation is believed to exist between two samples, but it is not known what items are paired or on what principle the correlation depends, it is often difficult to find the value of r_{12}.

When there is no correlation between the two samples, *i.e.*, when both of the samples are random or simple and so are independent, $r_{12} = 0$, and formula (136) reduces to

$$\epsilon_D{}^2 = \epsilon_1{}^2 + \epsilon_2{}^2. \tag{137}$$

Formulas (136) and (137) make no assumptions in addition to those involved in finding ϵ_1 and ϵ_2.

5. The Difference between Two Means.—A simpler formula than (136) for testing the *difference between the means of two matched samples* is

$$\epsilon_D = \frac{\sigma_d}{\sqrt{N}}, \tag{138}$$

where σ_d is the standard deviation of the differences between the paired values, and N is the number of pairs. The σ_d is estimated from the usual formula for the standard deviation. Formula (138) assumes that the experimental sample (*i.e.*, the random sample that receives the "treatment")[1] is a simple sample.

The scores of a simple sample of brothers and their sisters on a personality test are shown in Table 71. Are the means of the two series significantly different? Correlation is evidently present between brother and sister, so it is necessary to calculate the correlation between them or else to use formula (138). We shall do both, for comparison. For Table 71 we have, by formula (74),

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][(N\Sigma Y^2 - (\Sigma Y)^2]}},$$

$$r = \frac{40(1215) - 212(203)}{\sqrt{[40(1,320) - (212)^2][40(1,231) - (203)^2]}},$$

$$r = .70.$$

Also,

$$\sigma_d = \sqrt{\left(\frac{\Sigma d^2}{N}\right) - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{\left(\frac{121}{40}\right) - \left(\frac{9}{40}\right)^2},$$

$$\sigma_d = 1.73.$$

Now, the standard error squared of the mean of the X's is, by formula (118) of Chap. XII,

$$\epsilon_{M_s}{}^2 = \frac{\sigma^2}{N},$$

$$\sigma^2 = \frac{1,320}{40} - \left(\frac{212}{40}\right)^2 = 4.91,$$

$$\epsilon_{M_s}{}^2 = \frac{4.91}{40} = 0.12.$$

[1] See Chaps. III and IV.

TABLE 71.—PERSONALITY TEST SCORES OF 40 PAIRS OF BROTHERS AND
SISTERS. (HYPOTHETICAL DATA)

Brother (X)	Sister (Y)	XY	X²	Y²	d	d²
8	5	40	64	25	3	9 .
3	4	12	9	16	−1	1
8	7	56	64	49	1	1
2	2	4	4	4	0	0
2	3	6	4	9	−1	1
4	7	28	16	49	−3	9
6	5	30	36	25	1	1
5	3	15	25	9	2	4
7	9	63	49	81	−2	4
10	9	90	100	81	1	1
10	8	80	100	64	2	4
1	3	3	1	9	−2	4
9	7	63	81	49	2	4
8	6	48	64	36	2	4
7	5	35	49	25	2	4
7	8	56	49	64	−1	1
5	4	20	25	16	1	1
5	6	30	25	36	−1	1
5	5	25	25	25	0	0
4	6	24	16	36	−2	4
4	4	16	16	16	0	0
4	3	12	16	9	1	1
4	1	4	16	1	3	9
3	3	9	9	9	0	0
3	2	6	9	4	1	1
3	5	15	9	25	−2	4
3	4	12	9	16	−1	1
5	7	35	25	49	−2	4
8	10	80	64	100	−2	4
6	5	30	36	25	1	1
6	4	24	36	16	2	4
7	7	49	49	49	0	0
7	4	28	49	16	3	9
5	8	40	25	64	−3	9
4	1	4	16	1	3	9
2	2	4	4	4	0	0
6	5	30	36	25	1	1
4	3	12	16	9	1	1
7	6	42	49	36	1	1
5	7	35	25	49	−2	4
212	203	1,215	1,320	1,231	9	121

Similarly, for the Y's,

$$\epsilon_{M_y}^2 = \frac{\sigma_Y^2}{N},$$

$$\sigma_Y^2 = \frac{1{,}231}{40} - \left(\frac{203}{40}\right)^2 = 5.02,$$

$$\epsilon_{M_y}^2 = \frac{5.02}{40} = 0.13.$$

Substituting in formula (136),

$$\epsilon_D^2 = 0.12 - 2(0.7)(0.35)(0.36) + 0.13,$$
$$\epsilon_D^2 = 0.07, \qquad \text{or} \qquad \epsilon_D = 0.27.$$

Now,

$$M_x = \frac{\Sigma X}{N} = \frac{212}{40} = 5.30,$$

$$M_Y = \frac{\Sigma Y}{N} = \frac{203}{40} = 5.08,$$

$$C.R.^* = \frac{5.30 - 5.08}{0.27} = 0.81$$

The critical ratio is much less than two, so the difference between the two means is not significant.

Substituting next in formula (138),

$$\epsilon_D = \frac{1.73}{\sqrt{40}} = 0.27,$$

which quickly gives the same value obtained by the longer method.

The meaning of this result is that there is no more difference between the scores of brothers and sisters on a personality test than might be attributed to random errors of sampling.

Suppose we had neglected the correlation between the data of Table 71, and used formula (137) to test the significance of the difference between the two means. How much would the result have been changed? We have, from formula (137),

$$\epsilon_D^2 = 0.12 + 0.13 = 0.25$$

and

$$C.R. = \frac{5.30 - 5.08}{\sqrt{0.25}} = 0.44.$$

---

* In testing differences, the critical ratio ($C.R.$) is the ratio of the difference to the standard error of the difference.

The correction for dependence thus almost doubled the critical ratio, although in this instance it did not change the verdict regarding the significance of the difference.

When the hypothesis to be tested is that two simple samples were drawn from the same universe, the best estimate of any parameter is found by pooling the two samples. For example, if we are testing the difference between the means of two samples, formula (137) becomes

$$\epsilon_D{}^2 = \frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2} = \sigma^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right), \tag{139}$$

where N_1 is the number of cases in the first sample, N_2 is the number of cases in the second sample, and σ^2 is found from the two samples combined by the equation

$$\sigma^2 = \frac{\sum\limits^{N_1} X_1{}^2 - N_1 M^2{}_{X_1} + \sum\limits^{N_2} X_2{}^2 - N_2 M^2{}_{X_2}}{N_1 + N_2} \tag{140}[1]$$

where X_1 is any value of the variate in the first sample, M_{X_1} is the mean of the first sample, X_2 is any value of the variate in the second sample, and M_{X_2} is the mean of the second sample.

Table 72, below, gives the scores of 75 communities on a community organization test, the sample of communities being

TABLE 72.—SCORES OF 75 COMMUNITIES ON A COMMUNITY ORGANIZATION TEST

Score (X)	Communities (f)	d	fd	fd^2
80–99	7	2	14	28
60–79	15	1	15	15
40–59	29	0	0	0
20–39	13	−1	−13	13
0–19	11	−2	−22	44
Total........	75		− 6	100

simple and independent of the sample of 100 communities in Table 69 of Chap. XII. The mean score of Table 72 is 48.4, and its standard deviation is 23. Let us test the hypothesis

[1] This formula gives simply a weighted mean of the two variances, $\sigma_1{}^2$ and $\sigma_2{}^2$, and should not be confused with formula (29) of Chap. VIII, which gives the variance of combined distributions.

that these two tables represent simple samples from the same universe. The samples being independent by definition, we require formula (139). The variance of the two samples combined is found by formula (140), expressed in frequency form:

$$\sigma^2 = \frac{i^2\left[\sum_{}^{N_1} f_1 d_1{}^2 - \frac{\left(\sum_{}^{N_1} f_1 d_1\right)^2}{N_1} + \sum_{}^{N_2} f_2 d_2{}^2 - \frac{\left(\sum_{}^{N_1} f_2 d_2\right)^2}{N_2}\right]}{N_1 + N_2}, \quad (140a)$$

$$\sigma^2 = \frac{(20)^2\left[117 - \frac{(-7)^2}{100} + 100 - \frac{(-6)^2}{75}\right]}{100 + 75},$$

$$\sigma^2 = 493.78.$$

Substituting this value in formula (139),

$$\epsilon_D^2 = 493.78(\tfrac{1}{100} + \tfrac{1}{75}) = 11.52,$$
$$\epsilon_D = 3.394.$$

We therefore have

$$C.R. = \frac{(M_{x_1} - M_{x_2})}{\epsilon_D} = \frac{48.6 - 48.4}{3.39} = 0.0589.$$

Evidently, the data of the two tables might well represent simple samples from the same universe, as far as their mean values are concerned.

If it is believed that two simple samples are from different universes, and it is wanted to test whether the difference between their means falls within the range of chance error so that it might sometimes be obliterated by sampling error, formula (137) takes the form

$$\epsilon_D{}^2 = \frac{\sigma_1{}^2}{N_1} + \frac{\sigma_2{}^2}{N_2}. \quad (141)$$

Applying this formula to Tables 69 and 72, we get

$$\epsilon_D{}^2 = \frac{466.56}{100} + \frac{530.77}{75} = 11.74,$$
$$\epsilon_D = 3.43,$$

which is slightly larger than the standard error obtained on the assumption that the samples are from the same universe. Since

the critical ratio is only

$$C.R. = \frac{48.6 - 48.4}{3.43} = .0583,$$

we interpret it to mean that if there is a real difference between the universes from which the two samples came, it may easily be reduced to zero or nearly zero in random samples.

Sometimes two simple samples are taken from the same universe, and the mean of sample 1 is tested against the mean of the two samples combined. Correlation is thereby introduced, and the appropriate formula is then

$$\epsilon_D{}^2 = \frac{\sigma^2 N_2}{N_1(N_1 + N_2)}. \tag{142}$$

where σ^2 is found from the two samples combined, using formula (140) or formula (140a). This formula leads to the same critical ratio as formula (139). To show this, let us test the mean score (48.4) of the 75 communities in Table 72 against the mean score (48.51) of Table 72 and Table 69 of Chap. XII combined, on the theory that the two samples together give a better picture of the universe of communities from which the samples were drawn than does either one alone. Substituting in formula (142) the values previously found,

$$\epsilon_D{}^2 = \frac{493.78(100)}{75(75 + 100)},$$
$$\epsilon_D{}^2 = 3.76,$$
$$\epsilon_D = 1.9396,$$
$$C.R. = \frac{(48.51 - 48.4)}{1.9396} = 0.0589,$$

which is identical with the result previously obtained.

6. The Difference between Two Proportions.—Although we have so far dealt only with the differences between means of samples, the same types of formula hold for other statistics. For example, if we are testing the difference between two *proportions*, the formulas corresponding to formulas (139), (140), (141), and (142) are, in order:

Two simple samples from the same universe,

$$\epsilon_D{}^2 = \bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \tag{143}$$

where

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}. \tag{144}$$

Two simple samples from different universes,

$$\epsilon_D{}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}. \tag{145}$$

Two simple samples from the same universe, the proportion of successes in the first tested against the mean proportion of successes in the two combined,

$$\epsilon_D{}^2 = \frac{\bar{p}\bar{q}n_2}{(n_1 + n_2)n_1}. \tag{146}$$

Can samples 1 and 2 of Table 73, below, be simple samples from the same universe, in the proportion of families having seven or more members? Applying formula (143), we need the value of $\bar{p}$ from formula (144):

$$\bar{p} = \frac{(132)(\frac{14}{132}) + (134)(\frac{9}{134})}{132 + 134} = .0865,$$

$$\bar{q} = 1.0000 - .0865 = .9135,$$

Whence

$$\epsilon_D{}^2 = (.0865)(.9135)(\tfrac{1}{132} + \tfrac{1}{134}) = .001188,$$

$$\epsilon_D = 0.0345,$$

$$C.R. = \frac{(\frac{14}{132} - \frac{9}{134})}{.0345} = 1.13.$$

TABLE 73.—TWO SAMPLES OF FAMILIES, CLASSIFIED BY NUMBER OF MEMBERS

Members in family	Sample 1 (f_1)	Sample 2 (f_2)	Total
1– 2	56	66	122
3– 4	40	42	82
5– 6	22	17	39
7– 8	6	5	11 ⎫
9–10	4	1	5 ⎪
11–12	2	2	4 ⎬
13–14	0	1	1 ⎪
15–16	2	0	2 ⎭
Total..........	132	134	266

From Appendix Table 1, we find a probability of about 26 in 100 that a difference greater than that observed between the proportions in the two samples might occur by sampling error, under the conditions assumed. If now we use the alternative formula (146), we get

$$\epsilon_D{}^2 = \frac{(.0865)(.9135)(134)}{(132 + 134)132} = .000302,$$
$$\epsilon_D = 0.0174,$$
$$C.R. = \frac{\left(\frac{14}{132} - \frac{23}{266}\right)}{.0174} = 1.13.$$

Notice again that this is the same critical ratio as that obtained just above by the use of formula (143).

7. The Difference between Two Correlation Coefficients.—To test the significance of the difference between two *correlation coefficients* from simple samples,[1] the variates being normally distributed and independent, it is necessary to convert r_1 and r_2 to z_1 and z_2, respectively. This is readily done by means of Appendix Table 5. The standard error of z is then found from the formula

$$\epsilon_{z_i} = \frac{1}{\sqrt{N - 3}}. \tag{147}$$

Suppose for the correlation between the linguistic ability and leadership scores of a group of children, we find $r_1 = .50$ from sample 1, and $r_2 = .60$ from sample 2, where $N_1 = N_2 = 50$. Is the difference between the two r's significant, or is it merely an accident of sampling? From Appendix Table 5, we find for $r_1 = .50$, $z_1 = .549$, and for $r_2 = .60$, $z_2 = .693$. By formula (147) we calculate the standard error of z,

$$\epsilon_{z_1} = \frac{1}{\sqrt{50 - 3}} = 0.146.$$

Hence the standard error of the difference $z_2 - z_1$ is, by formula (137),

$$\epsilon_D{}^2 = (.146)^2 + (.146)^2 = 0.0426.$$

So that

$$C.R. = \frac{.693 - .549}{\sqrt{.0426}} = 0.699.$$

[1] Of any size.

Since this critical ratio is well under two standard errors, we infer that the difference between the two r's is not significant.

8. Testing the Significance of a Sum.—The basic formulas (136) and (137) are the same for the sum as for the difference between two statistics, except that for the sum all the signs of formula (136) are positive.

9. Testing the Hypothesis of Simple Sampling.—Suppose we ask if Table 72 above can be a simple sample from a universe of communities in which the mean score is 40. We have $\epsilon_M = \sigma^*/\sqrt{N} = 23/\sqrt{75} = 2.66$, so that

$$C.R. = (48.4 - 40.0)/2.66 = 3.16.$$

Since the critical ratio is greater than two standard errors, it is not likely that the sample is a simple sample from a universe of communities whose mean score is 40. There are several possible explanations: (1) Table 72 may be a simple sample with an extreme mean that might rarely be drawn by chance from the given universe; (2) it may be a sample from the given universe, but not taken as a simple sample should have been; or (3) it may not be a sample from the given universe at all. There is no way to determine which of these possibilities is correct, unless it can be learned how the sample was actually taken.

The purpose of testing the difference between the two means 'in Sec. 5, above, might have been to discover whether or not they could occur in two simple samples from the same universe. The very low critical ratio of 0.0589 suggests an affirmative answer, but it cannot completely establish the fact. For example, the low critical ratio might be due to the small size of the samples or to the presence of correlation, or it might be an accident not connected with random sampling.

The same test can, of course, be employed to determine whether a sample might be random or Poisson, by merely using the standard error formula appropriate in each case.

10. The Significance of the Difference between Two or More Frequency Distributions.—A more complete test of the hypothesis that two or more samples are simple samples from the same universe is possible by the Chi-square method, which goes beyond the comparison of single statistics (*e.g.*, means) to the comparison of whole distributions. In Table 73 we have two

* σ found from formula (118), Chap. XII.

samples of families distributed according to number of members. We can use either formula (148) or formula (149) to find Chi square (χ^2). Formula (148) is applicable only to the case of two samples, when the total χ^2 for each row is not wanted, but is quicker than formula (149). We shall apply it here.

$$\chi^2 = \frac{1}{\bar{p}\bar{q}}[\Sigma(_rf_1 \cdot _rp_1) - n_1\bar{p}], \qquad (148)$$

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}, \qquad (149)$$

where

$$\bar{p} = \frac{n_1}{N},$$

$$_rp_1 = \frac{_rf_1}{_rn},$$

$$\bar{q} = 1 - \bar{p},$$

$$f_t = \frac{_rn n_c}{N}.$$

$_rf_1$ is the frequency in row r of col. 1, n_c is the total of any column, c, n_1 is the total of col. 1, $_rn$ is the total of any row r, f_o is any observed frequency, f_t is any theoretical or expected frequency, and N is the total of the whole table. For Table 73 we have

$$\bar{p} = \tfrac{132}{266} = 0.496,$$
$$\bar{q} = 1 - 0.496 = 0.504,$$
$$_1p_1 = \tfrac{56}{122} = 0.459,$$
$$_2p_1 = \tfrac{40}{82} = 0.488,$$
$$_3p_1 = \tfrac{22}{39} = 0.564,$$
$$_4p_1 = \tfrac{14}{23} = 0.609.$$

To get $_4p_1$, the frequencies of the last five rows were combined, in accordance with the rule that no cell should contain less than five expected frequencies. Substituting in formula (148),

$$\chi^2 = \frac{1}{.496(.504)}[56(.459) + 40(.488) + 22(.564) + 14(.609)$$
$$- 132(.496)],$$

$$\chi^2 = 2.76.$$

Entering a table of Chi square (Appendix Table 2) with

$$r - 1 = 4 - 1 = 3 \text{ degrees of freedom}$$

(*i.e.*, one less than the number of rows, counting the five combined rows as one), we find a probability P between .30 and .50 that the differences between the two samples might be due to errors of simple sampling from the same universe. The test therefore fails to show that the two sample distributions of families differ significantly in number of members.

The Chi-square test may also be used to investigate whether a sample is a simple sample from a known universe, if the distribution of the universe is known. The universe distribution, with N equated to that of the sample, simply takes the place of one of the samples in Table 73.

To test whether more than two samples are from the same universe, it is necessary to find Chi square by formula (149). Its application to three samples in Table 74 is shown below.

TABLE 74.—THREE SAMPLES OF FAMILIES, CLASSIFIED BY NUMBER OF MEMBERS

Members in family	Sample 1			Sample 2			Sample 3			Total
	f_o	f_t	$\frac{(f_o - f_t)^2}{f_t}$	f_o	f_t	$\frac{(f_o - f_t)^2}{f_t}$	f_o	f_t	$\frac{(f_o - f_t)^2}{f_t}$	
1– 2	56	56.34	.00205	66	57.20	1.35385	53	61.46	1.16452	175
3– 4	40	40.57	.00801	42	41.18	.01633	44	44.25	.00141	126
5– 6	22	21.57	.00857	17	21.90	1.09635	28	23.53	.84917	67
7– 8	6	13.52	.01704	5	13.73	1.62949	9	14.75	1.22457	20
9–10	4	......	....	1	......	....	3	......	....	8
11–12	2	......	....	2	......	....	5	......	....	9
13–14	0	......	....	1	......	....	1	......	....	2
15–16	2	......	....	0	......	....	1	......	....	3
Total.......	132	132.01	0.03567	134	134.00	4.09602	144	143.99	3.23967	410

The expected frequency, f_t, in any cell is found, as explained in Chap. IX, by dividing the table total into the product of the row and column totals. For example, the expected frequency in the class interval 3–4, sample 2, is $126(134)/410 = 41.18$; in the class interval 5–6, sample 3, it is $67(144)/410 = 23.53$; and so on. The last five rows are combined, because four of them have fewer than five expected frequencies. After combining, the expected frequencies $(132)(42)/410 = 13.52$. By formula (149),

$$\chi^2 = 7.37136.$$

The degrees of freedom are

$$(c - 1)(r - 1) = (3 - 1)(4 - 1) = 6,$$

where c is the number of columns, and r is the number of rows, counting all combined rows as one. Entering a table of χ^2 with six degrees of freedom, we see that $\chi^2 = 12.59$ for $P = .05$. That is, a value of χ^2 as large as 12.59 may be expected by chance five times in 100. Smaller values of χ^2 will, of course, occur more often by chance. Since our value of χ^2 is only 7.37136, it cannot be regarded as significant. The test therefore furnishes no evidence that our three samples are not simple samples from the same universe.

11. The Significance of the Difference between Statistics from More than Two Samples.—In testing the significance of the differences between statistics (*e.g.*, means) from three or more samples, the probability of finding a significant difference by chance is greater than in the case of only one difference, just as the probability of getting an ace at cards is greater when we draw twice from the deck than when we draw only once.

A formula that takes this into account is the following:

$$C.R. = \sqrt{\frac{n - 1}{n^2}} \sum_{}^{n} \frac{d_i}{\epsilon_i}, \qquad (150)$$

where d_i is the difference between any two independent statistics, ϵ_i is its standard error, and n is the number of differences. In

TABLE 75.—SIX SAMPLES OF 50 JUVENILE DELINQUENTS EACH, AND SIX CONTROL SAMPLES, SHOWING PERCENTAGES NEUROTIC

Samples	Percentage delinquents neurotic	Percentage nondelin- quents neurotic	d_i	ϵ_i	$\dfrac{d_i}{\epsilon_i}$
1 and 1a..............	4	6	−2	4.4	−0.45
2 and 2a..............	10	4	6	5.1	1.18
3 and 3a..............	2	4	−2	3.4	−0.59
4 and 4a..............	0	2	−2	2.0	−1.00
5 and 5a..............	4	8	−4	4.7	−0.85
6 and 6a..............	6	2	4	3.9	1.03
Total..............					−0.68

Table 75 we have six simple samples of juvenile delinquents and six simple samples of nondelinquents, each containing 50 boys. Using formula (143) to find the standard error of the six differences, we have, for the first pair of samples in the table,

$$\epsilon_1 = \sqrt{\bar{p}\bar{q}\left(\frac{2}{n}\right)} = \sqrt{5(95)\left(\frac{2}{50}\right)} = 4.4$$

The standard errors of the other differences are found similarly, and entered in Table 75. Substituting from the last column of the table in formula (150),

$$C.R. = \sqrt{\frac{6-1}{(6)^2}}\,(-0.68) = -0.25.$$

From a table of normal areas (Appendix Table 1), it appears that a positive or negative critical ratio greater than this might occur by chance over 80 times in 100 trials, so that there is no evidence that the samples of delinquents differ significantly from the samples of nondelinquents in respect to the percentages neurotic.

Formula (150) is applicable to any set of independent critical ratios, including those from random or Poisson samples, if random or Poisson standard error formulas are used to find the values of ϵ_i.

Exercises

1. Correlate the birth rates of Table 70 of Chap. XII with the fertility ratios of the same sample counties in the table of Exercise 11 in Chap. XII, and test whether the correlation coefficient is significantly greater than zero.

2. In the table of Exercise 3, below, test the hypothesis that rural farm and rural nonfarm families (1) are from the same universe in respect to mean size of family; (2) are from different universes, but their means might sometimes be approximately equal as a result of sampling error.

3. In the table below test the assumption that urban families might be a random sample from a universe which is best represented by the three samples combined. Use the mean as the criterion.

SAMPLE OF 2,992 FAMILIES BY SIZE, FOR URBAN AND RURAL AREAS, UNITED
STATES, 1930

Members in family	Urban families	Rural farm families	Rural nonfarm families	Total
1..................	140	34	62	236
2..................	436	121	141	698
3..................	384	119	120	623
4..................	315	110	99	524
5..................	202	88	68	358
6..................	118	66	43	227
7..................	66	47	27	140
8..................	37	32	16	85
9..................	20	20	9	49
10.................	10	12	5	27
11.................	5	6	2	13
12 or more*.......	4	6	2	12
Sample total.......	1,737	661	594	2,992
Universe total....	17,400,000	6,600,000	5,900,000	29,900,000

* Count as 13.

4. Do the matched delinquents and nondelinquents in the sample
below differ significantly in mean I.Q.?

INTELLIGENCE QUOTIENTS OF A RANDOM SAMPLE OF 25 MALE JUVENILE
DELINQUENTS AND 25 MALE NONDELINQUENTS, MATCHED BY AGE,
FAMILY INCOME, AND PLACE OF RESIDENCE

Pair number	Intelligence quotient	
	Delinquent	Nondelinquent
1	103	99
2	80	92
3	114	106
4	100	104
5	91	88
6	73	80
7	105	109
8	98	94
9	86	90
10	101	97
11	92	89
12	86	91
13	93	90
14	90	97
15	79	84

INTELLIGENCE QUOTIENTS OF A RANDOM SAMPLE OF 25 MALE JUVENILE
DELINQUENTS AND 25 MALE NONDELINQUENTS, MATCHED BY AGE,
FAMILY INCOME, AND PLACE OF RESIDENCE.—(*Continued*)

Pair number	Intelligence quotient	
	Delinquent	Nondelinquent
16	108	96
17	82	91
18	95	86
19	74	83
20	102	97
21	105	99
22	97	103
23	88	91
24	94	84
25	99	106
Total........	2,335	2,346

5. Combining rural farm and rural nonfarm families in the table of Exercise 3, above, is there a significant difference between the mean size of family in urban and rural areas according to this sample taken proportionally at random from the two types of areas?

6. Apply Chi-square to the table of Exercise 3 above to test (1) the hypothesis that rural farm and rural nonfarm families are simple samples from the same universe; (2) the hypothesis that urban, rural farm, and rural nonfarm families are simple samples from the same universe.

7. In the table of Exercise 3 above, test the hypothesis that the means of the three simple samples are from the same universe.

8. Test the hypothesis that families of odd sizes and families of even sizes in the table of Exercise 3 above are Poisson samples from the same stratified universe.

9. Test the hypothesis that the value of a selected statistic from each of the samples drawn in Exercise 4 of Chap. XII does not differ significantly from the known value of the corresponding parameter in the universe.

10. Test the hypothesis that the value of a selected statistic from the sample drawn in Exercise 5 of Chap. XII does not differ significantly from the known value of the corresponding parameter in the universe.

References

Same as for Chap. XII.

CHAPTER XIV

TIME SERIES ANALYSIS

Values of a variable (*e.g.*, infant death rates) given at successive intervals of time (*e.g.*, yearly) form a *time series*. Such series are especially important in economics and are also necessary in the study of vital statistics,[1] of trends in public expenditures for relief, and many other topics. This chapter describes methods for their analysis.

As an illustrative problem, let us inquire what the state of Wisconsin has accomplished in reducing infant mortality. Figures giving deaths per 1,000 live births from 1908 through 1935 are shown as a time series in Table 76.

TABLE 76.—WISCONSIN INFANT MORTALITY RATES, 1908–1935*

Year	Infant deaths per 1,000 live births	Year	Infant deaths per 1,000 live births
1908	107	1922	70
1909	120	1923	70
1910	109	1924	64
1911	103	1925	67
1912	95	1926	67
1913	97	1927	60
1914	83	1928	61
1915	78	1929	60
1916	86	1930	56
1917	78	1931	53
1918	79	1932	50
1919	79	1933	48
1920	77	1934	49
1921	72	1935	46

* *Report of the Wisconsin Bureau of Vital Statistics*, 1934–1935, p. 284.

The first step that is usually taken in time series analysis is to plot the data. This is done for Table 76 in the lower graph of

[1] Birth rate, death rates, marriage rates, and so on.

Fig. 55. Examination of this figure shows a striking decline in infant mortality in Wisconsin over a 28-year period.
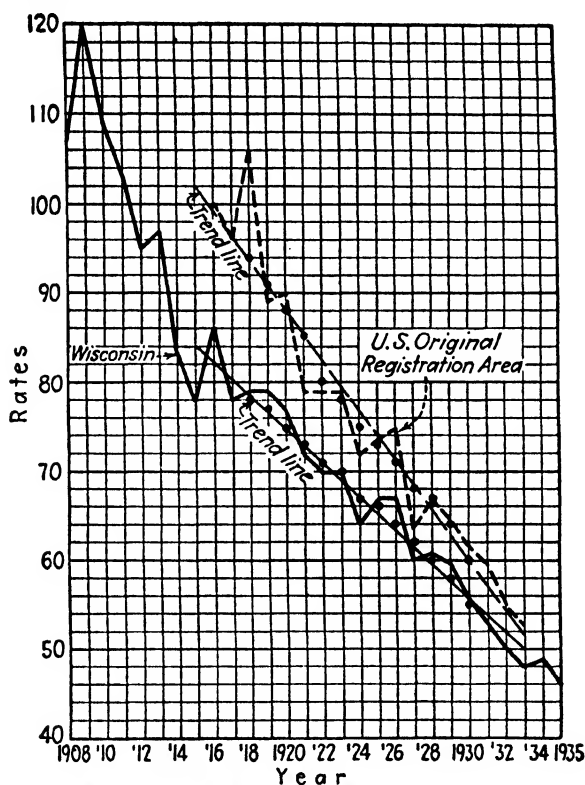
Note: Dots indicate moving averages

Fig. 55.—Infant mortality rates for Wisconsin and for the United States, 1908–1935. (From Tables 76, 77, and 80.)

1. The Secular Trend: A Straight Line.—Suppose, further, that we want to compare the infant mortality record in Wisconsin with that of other states in the United States. Data for the original birth registration area of 10 states and the District of Columbia[1] are available for the period 1915 through 1933, and are entered in Table 77. They are plotted as a dotted line in Fig. 55. It is seen, from this figure, that infant mortality has been less in Wisconsin than in the original registration area

[1] Connecticut, Maine, Massachusetts, Michigan, New Hampshire, New York, Pennsylvania, Rhode Island, Vermont, and the District of Columbia.

throughout the entire period of comparison. But has the *rate of decline* since 1915 been greater in Wisconsin or in the original registration area?

TABLE 77.—INFANT MORTALITY IN THE ORIGINAL BIRTH REGISTRATION AREA OF THE UNITED STATES, 1915–1933*

Year	Infant deaths per 1,000 live births	Year	Infant deaths per 1,000 live births
1915	100	1925	74
1916	100	1926	75
1917	96	1927	64
1918	106	1928	67
1919	89	1929	65
1920	90	1930	62
1921	79	1931	60
1922	79	1932	55
1923	79	1933	53
1924	72		

* From *Birth, Stillbirth, and Infant Mortality Statistics,* 1933, p. 7, U. S. Bureau of the Census.

The answer is to determine which of the two series has the steeper slope. Inspection shows that both graphs are irregular and saw-toothed in shape, so that the slope sometimes of one and sometimes of the other is the steeper. What we must do is to remove the irregularities in the two series, *i.e.*, reduce them to smooth curves. To do this is to find the *secular trend*, meaning the general direction of the series over a considerable period of time, freed from confusing oscillations. To answer our particular question in the present case, it seems appropriate to fit straight lines to the data, since more complex smooth curves do not describe the declining death rates any better.

We have already learned to fit a straight line by the device of least squares, in determining the regression equation in linear correlation. The normal equations for finding the values of a and b in the line of best fit are

$$b = \frac{\Sigma x Y}{\Sigma x^2}, \tag{151}$$

$$a = M_y, \tag{152}$$

where x is a deviate from the midyear of an *odd*[1] number of years, Y is the infant death rate, and the origin is at the midyear or mean of the X's. The values of a and b so found are substituted in the equation of the straight line.

$$Y_o = a + bx. \tag{153}$$

We set up Table 78 to obtain these values for the Wisconsin series, and Table 79 for the registration area series.

TABLE 78.—FITTING A STRAIGHT LINE TO THE WISCONSIN DATA OF TABLE 76

Year	Year (x_1)	Infant death rate (Y)	x_1Y	x_1^2
1915	−9	78	−702	81
1916	−8	86	−688	64
1917	−7	78	−546	49
1918	−6	79	−474	36
1919	−5	79	−395	25
1920	−4	77	−308	16
1921	−3	72	−216	9
1922	−2	70	−140	4
1923	−1	70	− 70	1
1924	0	64	0	0
1925	1	67	67	1
1926	2	67	134	4
1927	3	60	180	9
1928	4	61	244	16
1929	5	60	300	25
1930	6	56	336	36
1931	7	53	371	49
1932	8	50	400	64
1933	9	48	432	81
		$\Sigma Y = 1,275$ $M_y = 67.1$	$\Sigma x_1Y = -1,075$	$\Sigma x_1^2 = 570$

From Table 78,

$$b_1 = \frac{-1,075}{570} = -1.886,$$

[1] If the series includes an even number of years, one of them may be dropped to give an odd number, so that the convenient short formulas (151) and (152) may be used, instead of the more laborious normal equations for a straight line given in Chap. X.

and

$$a_1 = 67.1,$$

so that, approximately,

$$Y_1 = 67 - 1.89x_1 \qquad (153a)$$

is the equation of the straight-line secular trend fitted to infant mortality rates in Wisconsin with origin at 1924.

Similarly, from Table 79, we find the equation of the straight-line trend through infant mortality rates in the original registration area of the United States to be approximately

$$Y_2 = 77 - 2.75x_2. \qquad (153b)$$

TABLE 79.—FITTING A STRAIGHT LINE TO THE ORIGINAL REGISTRATION AREA DATA OF TABLE 77

Year	Year (x_2)	Infant death rate (Y)	x_2Y	x_2^2
1915	−9	100	−900	(see Table 78)
1916	−8	100	−800	
1917	−7	96	−672	
1918	−6	106	−636	
1919	−5	89	−445	
1920	−4	90	−360	
1921	−3	79	−237	
1922	−2	79	−158	
1923	−1	79	− 79	
1924	0	72	0	
1925	1	74	74	
1926	2	75	150	
1927	3	64	192	
1928	4	67	268	
1929	5	65	325	
1930	6	62	372	
1931	7	60	420	
1932	8	55	440	
1933	9	53	447	
		$\Sigma Y = 1,465$ $M_y = 77.1$	$\Sigma x_2Y = -1,569$	$\Sigma x_2^2 = 570$

We are now in a position to answer the question, Does the trend line for Wisconsin or that for the original registration area

have the steeper slope? We see that the slope of the Wisconsin line is $b_1 = -1.89$, whereas the slope of the original registration area line is $b_2 = -2.75$. The negative signs mean that as x increases, *i.e.*, as time passes, Y, the infant death rate, decreases. Evidently, the trend of infant mortality has been decreasing $\dfrac{2.75}{1.89} = 1.5$ times as fast in the original registration area as in Wisconsin. The two lines of trend are plotted in Fig. 55 by substituting appropriate values of x in equations (153*a*) and (153*b*). For example, the ordinate of the line through the Wisconsin data is, if $x_1 = -9$, $Y_1 = 67 - 1.89 (-9) = 84$; and if $x_1 = 9$, $Y_1 = 67 - 1.89 (9) = 50$; so that the line is drawn through the points $(-9, 84)$ and $(9, 50)$.

In terms of percentages, the infant mortality rate declined an average of 2.13 per cent per year in Wisconsin, as compared with 2.56 per cent in the total registration area.

2. The Secular Trend : A Moving Average.—It is an important principle that any line or curve used to represent the secular trend of a series should be rather simple in form—a straight line if that is at all reasonable, otherwise seldom anything more complex than a second degree parabola ($Y = a + bX + cX^2$). The reasons are that a trend line that follows the original data too closely includes cyclical variations from which the secular trend should be freed, and it also fails to fulfill the primary purpose of a trend line, which is to show clearly the general direction, up or down, in which the series is moving.

Of course, a straight line may be a very poor fit for some series, so that if we want to generalize the trend without doing too much violence to the data we may need to fit another kind of curve, say, a parabola. Although the formulas differ, the general principles are the same.

A second method of determining secular trend, which usually allows the trend line to follow the original data more closely than a straight line does, should be explained. This is the method of the moving average, which is shown in Table 80. It is again preferable to average an odd number of years, because the results can then be more conveniently centered at a given year in the series. If cycles appear in the original series, the length of the moving average should be equal to the average period of a cycle from peak to peak, or some multiple thereof, if

the purpose is to represent the secular trend.   But if the moving average is used only to smooth out random fluctuations, its length should be less than that of an average cycle period. The shorter the period of the moving average, the more flexible is the resulting curve.   Inspection of Fig. 55 suggests the presence of possible cycles of about seven years in length in both series.   Accordingly, moving averages of seven years are shown in Table 80.

TABLE 80.—SEVEN-YEAR MOVING AVERAGES OF INFANT MORTALITY RATES IN WISCONSIN AND IN THE ORIGINAL REGISTRATION AREA OF THE UNITED STATES, 1915–1933

Year	Mortality rates		Seven-year moving averages	
	Wisconsin	Registration area	Wisconsin	Registration area
1915	78	100		
1916	86	100		
1917	78	96		
1918	79	106	78	94
1919	79	89	77	91
1920	77	90	75	88
1921	72	79	73	85
1922	70	79	71	80
1923	70	79	70	78
1924	64	72	67	75
1925	67	74	66	73
1926	67	75	64	71
1927	60	64	62	68
1928	61	67	61	67
1929	60	65	58	64
1930	56	62	55	61
1931	53	60		
1932	50	55		
1933	48	53		

The method is simply to add the first seven values of the series, and divide by 7.   Thus, for the Wisconsin series, we have $(78 + 86 + 78 + 79 + 79 + 77 + 72 = 549)\frac{1}{7} = 78.4$.   Then the first value in the table, 78, is dropped, and the eighth value,

70, is added, and again the sum is divided by 7:

$$(549 - 78 + 70)\tfrac{1}{7} = \tfrac{541}{7} = 77.3;$$

and so on.

Notice that a disadvantage of the moving average is that it reduces the length of the series by one less than the number of years averaged, or in this case $7 - 1 = 6$ years. When the moving averages are plotted as large dots in Fig. 55, it is seen that they give trend lines that agree very closely with the straight lines of best fit, especially in the case of the Wisconsin data.

It is helpful in selecting a secular trend to note that "if the actual data fall consistently above or below a line of trend for a considerable period, it is probable that the fit is not good."[1] This is not the case in Fig. 55.

3. Short-term Cycles.—The cycles in the Wisconsin and registration area series may be shown more clearly than in Fig. 55
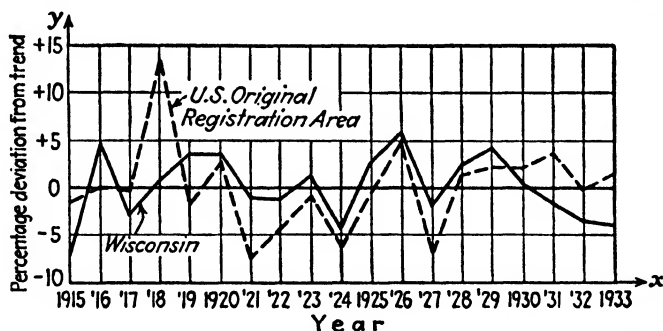
FIG. 56.—Infant mortality rates in Wisconsin and the original registration area of the United States, 1915–1933: cyclical deviations from linear trends. (From Table 82.)

by expressing the original rates as percentages of the trend, using for the latter either the values lying on the straight lines of best fit or the moving averages just found. If we choose the former, the results are shown in the last two columns of Table 81. Thus, from Table 80, for 1915, we have 78, and from Table 81, 84.01, so that $100(78/84.01) = 92.85$. Any cyclical tendencies in these percentages of trend will stand out even more if we subtract 100 per cent from each of them, thus expressing them as positive

[1] F. C. MILLS, *Statistical Methods*, p. 290, Henry Holt and Company, Inc., New York, 1924.

and minus deviations.   This is done in Table 82,[1] and the
resulting cyclical deviations are plotted in Fig. 56.

From Fig. 56 it appears that only short and erratic cycles
occur in infant mortality rates in Wisconsin and in the original

TABLE 81.—INFANT MORTALITY RATES IN WISCONSIN AND IN THE ORIGINAL
REGISTRATION AREA OF THE UNITED STATES, 1915–1933
Straight-line Trend Values and Observed Values as Percentages of the
Trend Values

Year	Linear trend values		Observed rates as per cent of trend	
	Wisconsin	Registration area	Wisconsin	Registration area
1915	84.01	101.75	92.85	98.28
1916	82.12	99.00	104.72	101.01
1917	80.23	96.25	97.22	99.74
1918	78.34	93.50	100.84	113.37
1919	76.45	90.75	103.34	98.07
1920	74.56	88.00	103.27	102.27
1921	72.67	85.25	99.08	92.67
1922	70.78	82.50	98.90	95.76
1923	68.89	79.75	101.61	99.06
1924	67.00	77.00	95.52	93.51
1925	65.11	74.25	102.90	99.66
1926	63.22	71.50	105.98	104.90
1927	61.33	68.75	97.83	93.09
1928	59.44	66.00	102.62	101.52
1929	57.55	63.25	104.26	102.77
1930	55.66	60.50	100.61	102.48
1931	53.77	57.75	98.57	103.90
1932	51.88	55.00	96.38	100.00
1933	49.99	52.25	96.02	101.44

registration area over the period 1915 through 1933.   Slightly
different results would have been obtained if the moving average
instead of the straight line had been used as the index of trend.

[1] Notice that the first two columns of Table 82 should each sum to zero.
They fail to do so because we disregarded decimals in the equations of the
lines of best fit.

TABLE 82.—INFANT MORTALITY RATES IN WISCONSIN AND THE ORIGINAL
REGISTRATION AREA OF THE UNITED STATES, 1915–1933
Percentage Deviations from Straight-line Trends

Year	Percentage deviations from trend		x'^2	y'^2	$(x'y')$
	Wisconsin (x')	Registration area (y')			
1915	−7.15	− 1.72	51.12	2.96	12.30
1916	+4.72	+ 1.01	22.28	1.02	4.77
1917	−2.78	− 0.26	7.73	.07	0.72
1918	+0.84	+13.37	.71	178.76	11.23
1919	+3.34	− 1.93	11.16	3.72	− 6.45
1920	+3.27	+ 2.27	10.69	5.15	7.42
1921	−0.92	− 7.33	.85	53.73	6.74
1922	−1.10	− 4.24	1.21	17.98	4.66
1923	+1.61	− 0.94	2.59	.88	− 1.51
1924	−4.48	− 6.49	20.07	42.12	29.08
1925	+2.90	− 0.34	8.41	.12	− 0.99
1926	+5.98	+ 4.90	35.76	24.01	29.30
1927	−2.17	− 6.91	4.71	47.75	14.99
1928	+2.62	+ 1.52	6.86	2.31	3.98
1929	+4.26	+ 2.77	18.15	7.67	11.80
1930	+0.61	+ 2.48	.37	6.15	1.51
1931	−1.43	+ 3.90	2.04	15.21	− 5.58
1932	−3.62	0.00	13.10	0.00	0.00
1933	−3.98	+ 1.44	15.84	2.07	− 5.73
Total.........	+2.52	+ 3.50	233.65	411.68	118.26

To compare the amounts of fluctuation of the two series
around the line of trend, the percentage deviations of Table 82
are squared and summed, giving for the Wisconsin series,

$$\sigma_{x'} = \sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2} = \sqrt{\frac{233.65}{19} - \left(\frac{2.52}{19}\right)^2} = 3.51,$$

and for the original registration area,

$$\sigma_{y'} = \sqrt{\frac{411.68}{19} - \left(\frac{3.50}{19}\right)^2} = 4.65.$$

We therefore conclude that the original registration area series is 1.32 times as variable as the Wisconsin series. Some of this difference is due to the abnormal rates of the war year 1918. We would expect such a result, as conditions affecting infant health are probably more variable over the whole registration area than in the single state of Wisconsin.

4. Correlation between the Short-term Cycles of Two Time Series.—Inspection of Fig. 56 shows that infant mortality rates tend to rise and fall together in Wisconsin and in the original registration area. This resemblance between the apparently erratic fluctuations of the two series may be symptomatic of the existence of general factors that produce cycles in infant deaths. The point is important enough to test with some care. We may ask, just how much relationship is there between the variations in infant mortality rates in Wisconsin and in the original registration area? To answer this question we need to know the value of the coefficient of correlation between the two time series, taking the deviations from the trend lines, as given in Table 82, instead of from the means of the series. It will be recalled that the formula for the Pearsonian coefficient of correlation is

$$r = \frac{\Sigma x'y' - NM_{x'}M_{y'}}{N\sigma_{x'}\sigma_{y'}}.$$

Taking the sum of the cross products, $\Sigma x'y' = 118.26$, from Table 82, $N = 19$ years from 1915 to 1933 inclusive, $\sigma_{x'} = 3.50$, and $\sigma_{y'} = 4.65$, as found above, we have

$$r = \frac{118.26 - 19\left(\frac{2.52}{19}\right)\left(\frac{3.50}{19}\right)}{19(3.51)(4.65)},$$

$$r = \frac{118.26 - 0.46}{301.11} = 0.39,$$

$$r^2 = 0.15.$$

So that the relationship between infant mortality rates in Wisconsin and in the original registration area from year to year enables us to predict one from a knowledge of the other only 15 per cent more accurately than if we judged one of the series from a knowledge of its own mean and variance.

Could it be that a correlation coefficient of $r = .39$ is due to random accidental correspondence between the cyclical fluctua-

tions in the two series? Although we are dealing here with two *historical* series, we have removed the secular trend, and this is sometimes regarded as warrant for applying the standard error to this situation. An inspection of the cycles in Fig. 56, however, suggests that some correlation between successive years still remains, so that we can hardly assume that the death rates in our series, regarded as a sample, are independent of one another. Under these conditions, the basic assumptions of simple sampling underlying the standard error formula which is appropriate in this case, *viz.*, $\epsilon_r = 1/\sqrt{N-1}$, are violated; so we are unable to answer the question asked at the beginning of the paragraph. However, the absence of much correlation between the Wisconsin rates and the original registration area rates suggests that the control of infant mortality is primarily a local problem. This should be further tested by comparing infant mortality rates in Wisconsin with those in adjoining states.

It is just as important to avoid the distorting effects of one or a few atypical, extreme values in correlating times series as in other correlation problems (see Chap. X, Table 49). For example, in Fig. 56 it appears that the war year 1918 was decidedly abnormal in its infant mortality rate, and the same is to some extent true of the depression year, 1933. In those two years there is much less agreement than usual between the two series. If we are interested primarily in knowing the amount of correlation between infant death rates in Wisconsin and in the registration area in normal years, it is, of course, desirable to omit the two atypical years from the computation of the correlation coefficient. This would make it necessary to fit a new trend line to the remaining 17 years of the series, and find the coefficient of correlation between the percentage deviations from it. In case we do not want to confine the investigation of the amount of association between the two series to "normal" years, which are not always easy to define objectively, and yet we do want to reduce the influence of the extreme or atypical values, it is probably advisable to resort to the coefficient of rank correlation. This coefficient, ρ, is calculated from Table 83, and has a value of .41.

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6(678)}{19(19^2 - 1)} = .41.$$

TABLE 83.—RANK OF OBSERVED RATES AS PER CENT OF TREND

Year	Wisconsin	Registration area	D	D²
1915	1	6	− 5	25
1916	18	11	7	49
1917	5	9	− 4	16
1918	11	19	− 8	64
1919	16	5	11	121
1920	15	14	1	1
1921	9	1	8	64
1922	8	4	4	16
1923	12	7	5	25
1924	2	3	− 1	1
1925	14	8	6	36
1926	19	18	1	1
1927	6	2	4	16
1928	13	13	0	0
1929	17	16	1	1
1930	10	15	5	25
1931	7	17	−10	100
1932	4	10	− 6	36
1933	3	12	9	81
Total........				678

As expected, the result of using ranks in this case is to increase the amount of correlation somewhat.

It often happens that the correlation of two time series is greater if one of them is lagged one or more years, so that the cycles correspond more closely. For example, if the marriage rate declines sharply, so does the birth rate, but not until about a year later. Therefore, to test the relationship between marriage and birth rates, the latter should be lagged by one year. That is, say, the 1930 birth rate should be paired with the 1929 marriage rate, etc. There is no indication that a lag is needed in correlating the two series with which we were dealing above.

5. Seasonal Fluctuations.—Data such as infant mortality rates may be obtained by months as well as by years. This affords an opportunity to study the seasonal fluctuations in infant deaths, *i.e.*, the variations in death rates that are associ-

TABLE 84.—INFANT MORTALITY RATES BY MONTHS, UNITED STATES
REGISTRATION AREA, 1928–1935*

Year	Month	Infant mortality rate per 1,000 live births in same month	Monthly trend rates	Observed rate as per cent of trend (2) ÷ (3) × 100	Monthly averages of observed rates as per cent of trend (Table 87)	Seasonal index	Cycles (4)−(6)
(1a)	(1b)	(2)	(3)	(4)	(5)	(6)	(7)
1928	Jan.	72.4	69.24	104.56	113.34	113.35	− 8.79
	Feb.	73.2	69.09	105.95	112.19	112.20	− 6.25
	Mar.	74.8	68.93	108.52	108.55	108.56	− .04
	Apr.	75.0	68.78	109.04	103.38	103.39	+ 5.65
	May	70.4	68.62	102.59	97.48	97.49	+ 5.10
	June	64.2	68.47	93.76	93.59	93.60	+ .16
	July	60.8	68.31	89.01	90.09	90.10	− 1.09
	Aug.	60.2	68.16	88.32	87.03	87.04	+ 1.28
	Sept.	63.4	68.00	93.24	91.20	91.21	+ 2.03
	Oct.	64.3	67.84	94.78	97.09	97.10	− 2.32
	Nov.	65.2	67.69	96.32	97.83	97.84	− 1.52
	Dec.	81.3	67.53	120.39	108.07	108.08	+12.31
1929	Jan	99.1	67.39	147.05	113.34	113.35	+33.70
	Feb.	84.8	67.22	126.15	112.19	112.20	+13.95
	Mar.	74.3	67.07	110.78	108.55	108.56	+ 2.22
	Apr.	66.1	66.91	98.79	103.38	103.39	− 4.60
	May	63.9	66.76	95.72	97.48	97.49	− 1.77
	June	57.8	66.60	86.79	93.59	93.60	− 6.81
	July	55.7	66.45	83.82	90.09	90.10	− 6.28
	Aug.	57.7	66.29	87.04	87.03	87.04	0.00
	Sept.	63.4	66.13	95.87	91.20	91.21	+ 4.66
	Oct.	64.9	65.98	98.36	97.09	97.10	+ 1.26
	Nov.	59.4	65.82	90.25	97.83	97.84	− 7.59
	Dec.	65.2	65.67	99.28	108.07	108.08	− 8.80
1930	Jan.	67.8	65.51	103.50	113.34	113.35	− 9.85
	Feb.	69.8	65.36	106.79	112.19	112.20	− 5.41
	Mar.	69.3	65.20	106.29	108.55	108.56	− 2.27
	Apr.	68.2	65.05	104.84	103.38	103.39	+ 1.45
	May	62.5	64.89	96.32	97.48	97.49	− 1.17
	June	61.4	64.74	94.84	93.59	93.60	+ 1.24
	July	59.3	64.58	91.82	90.09	90.10	+ 1.72
	Aug.	56.0	64.43	86.92	87.03	87.04	− .12
	Sept.	61.7	64.27	96.00	91.20	91.21	+ 4.79
	Oct.	67.1	64.12	104.65	97.09	97.10	+ 7.55
	Nov.	63.5	63.96	99.28	97.83	97.84	+ 1.44
	Dec.	69.8	63.81	109.39	108.07	108.08	+ 1.31
1931	Jan.	75.3	63.65	118.30	113.34	113.35	+ 4.95
	Feb.	74.6	63.49	117.50	112.19	112.20	+ 5.30
	Mar.	70.4	63.34	111.15	108.55	108.56	+ 2.59
	Apr.	65.7	63.18	103.99	103.38	103.39	+ .60
	May	56.4	63.03	89.48	97.48	97.49	− 8.01
	June	53.5	62.87	84.94	93.59	93.60	− 8.66
	July	54.1	62.72	86.26	90.09	90.10	− 3.84
	Aug.	54.3	62.56	86.80	87.03	87.04	− .24
	Sept.	58.5	62.41	93.73	91.20	91.21	+ 2.52
	Oct.	61.0	62.25	97.99	97.09	97.10	+ .89
	Nov.	58.1	62.10	93.56	97.83	97.84	− 4.28
	Dec.	57.3	61.94	92.51	108.07	108.08	−15.57
1932	Jan.	56.5	61.79	91.44	113.34	113.35	−21.91
	Feb.	57.5	61.63	93.30	112.19	112.20	−18.90
	Mar.	62.8	61.48	102.15	108.55	108.56	− 6.41
	Apr.	60.0	61.32	97.85	103.38	103.39	− 5.54
	May	57.8	61.17	94.49	97.48	97.49	− 3.00
	June	56.1	61.01	91.95	93.59	93.60	− 1.65
	July	55.2	60.85	90.71	90.09	90.10	+ .61
	Aug.	50.9	60.70	83.86	87.03	87.04	− 3.18
	Sept.	49.7	60.54	82.09	91.20	91.21	− 9.12
	Oct.	52.5	60.39	86.93	97.09	97.10	−10.17
	Nov.	60.4	60.23	100.28	97.83	97.84	+ 2.44
	Dec.	73.0	60.08	121.50	108.07	108.08	+13.42

* From *Births, Stillbirths, and Infant Mortality*, U. S. Bureau of the Census, annual publication.

TABLE 84.—INFANT MORTALITY RATES BY MONTHS, UNITED STATES REGISTRATION AREA, 1928–1935.*—(*Continued*)

Year	Month	Infant mortality rate per 1,000 live births in same month	Monthly trend rates	Observed rate as per cent of trend $(2)+(3)\times100$	Monthly averages of observed rates as per cent of trend (Table 87)	Seasonal index	Cycles (4)−(6)
(1a)	(1b)	(2)	(3)	(4)	(5)	(6)	(7)
1933	Jan.	71.2	59.92	118.83	113.34	113.35	+ 5.48
	Feb.	69.9	59.77	116.95	112.19	112.20	+ 4.75
	Mar.	60.1	59.61	100.82	108.55	108.56	− 7.74
	Apr.	56.3	59.46	94.69	103.38	103.39	− 8.70
	May	54.7	59.30	92.24	97.48	97.49	− 5.25
	June	56.2	59.15	95.01	93.59	93.60	+ 1.41
	July	51.9	58.99	87.98	90.09	90.10	− 2.12
	Aug.	50.1	58.84	85.15	87.03	87.04	− 1.89
	Sept.	54.9	58.68	93.56	91.20	91.21	+ 2.35
	Oct.	58.7	58.52	100.31	97.09	97.10	+ 3.21
	Nov.	58.0	58.37	99.37	97.83	97.84	+ 1.53
	Dec.	58.5	58.21	100.50	108.07	108.08	− 7.58
1934	Jan.	60.6	58.06	104.37	113.34	113.35	− 8.98
	Feb.	66.5	57.90	114.85	112.19	112.20	+ 2.65
	Mar.	67.7	57.75	117.23	108.55	108.56	+ 8.67
	Apr.	64.9	57.59	112.69	103.38	103.39	+ 9.30
	May	60.8	57.44	105.85	97.48	97.49	+ 8.36
	June	60.2	57.28	105.10	93.59	93.60	+11.50
	July	58.3	57.13	102.05	90.09	90.10	+11.95
	Aug.	52.2	56.97	91.63	87.03	87.04	+ 4.59
	Sept.	51.6	56.82	90.81	91.20	91.21	− .40
	Oct.	57.0	56.66	100.60	97.09	97.10	+ 3.50
	Nov.	59.3	56.51	104.94	97.83	97.84	+ 7.10
	Dec.	63.8	56.35	113.22	108.07	108.08	+ 5.14
1935	Jan.	66.7	56.19	118.70	113.34	113.35	+ 5.35
	Feb.	65.0	56.04	115.99	112.19	112.20	+ 3.79
	Mar.	62.3	55.88	111.49	108.55	108.56	+ 2.93
	Apr.	58.6	55.73	105.15	103.38	103.39	+ 1.76
	May	57.3	55.57	103.11	97.48	97.49	+ 5.62
	June	53.4	55.42	96.36	93.59	93.60	+ 2.76
	July	49.2	55.26	89.03	90.09	90.10	+ 1.07
	Aug.	47.7	55.11	86.55	87.03	87.04	− .49
	Sept.	46.3	54.95	84.26	91.20	91.21	− 6.95
	Oct.	51.0	54.80	93.07	97.09	97.10	− 4.03
	Nov.	53.9	54.64	98.65	97.83	97.84	+ .81
	Dec.	58.7	54.49	107.73	108.07	108.08	− .35

*From *Births, Stillbirths, and Infant Mortality*, U. S. Bureau of the Census, annual publication.

ated with spring, summer, fall, and winter. To do this, we must first separate the seasonal fluctuations from the secular trend, the short-term cycles, and the random fluctuations, all of which appear in the original monthly rates given in col. (2) of Table 84. We average the 12 monthly rates in each year in Table 84 to obtain annual rates, which are entered in Table 85, and plotted in Fig. 57. Inspection of Fig. 57 shows a decline in the infant mortality rate in five out of seven years, and suggests that a straight line probably is most appropriate to represent the secular trend. Table 85 shows the calculations needed to fit a linear trend to the annual rates by the method of least squares.
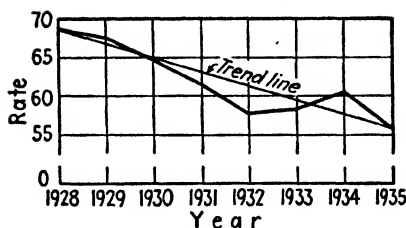
Fig. 57.—Annual infant mortality rates in the registration area of the United States, 1928–1933. (From Table 85.)

TABLE 85.—VALUES NEEDED FOR FITTING A STRAIGHT LINE TO THE ANNUAL INFANT MORTALITY RATES IN THE REGISTRATION AREA OF THE UNITED STATES, 1928–1935

Year	Infant death rate (Y)	Year code (X')	$X'Y$	X'^2	Trend values	Deviations from trend
1928	68.767	−3	−206.301	9	68.388	+ .379
1929	67.692	−2	−135.384	4	66.524	+1.168
1930	64.700	−1	− 64.700	1	64.660	+ .040
1931	61.592	0	0.000	0	62.796	−1.204
1932	57.700	+1	54.700	1	60.932	−3.232
1933	58.375	+2	116.750	4	59.068	− .693
1934	60.242	+3	180.726	9	57.204	+3.038
1935	55.842	+4	223.368	16	55.340	+ .502
Total.........	494.910		169.159	44		− .002
Mean.........	61.864	0.5				

Substituting the values found in Table 85 in the normal equations for determining the constants in the equation of a straight line, we have

$$b = \frac{\Sigma X'Y - NM_{x'}M_y}{\Sigma X'^2 - NM_{x'}^2}$$

$$a = M_y - bM_{x'},$$

$$b = \frac{169.159 - 8(0.5)(61.864)}{44 - 8(0.25)},$$

$$b = -1.864,$$

$$a = 61.864 - (-1.864)(0.5),$$

$$a = 62.796,$$

so that

$$Y_c = a + bX',$$

$$Y_c = 62.796 - 1.864X'. \tag{154}$$

From formula (154) the trend values shown in the next to the last column of Table 85 are estimated by substituting for X' its successive values taken from the third column of the table. The annual trend line is plotted in Fig. 57. The last column of Table 85, obtained by subtracting the trend values from the observed Y values, is inserted as a check on the arithmetic. Its sum is approximately zero, as it should be if the calculations are carried far enough.
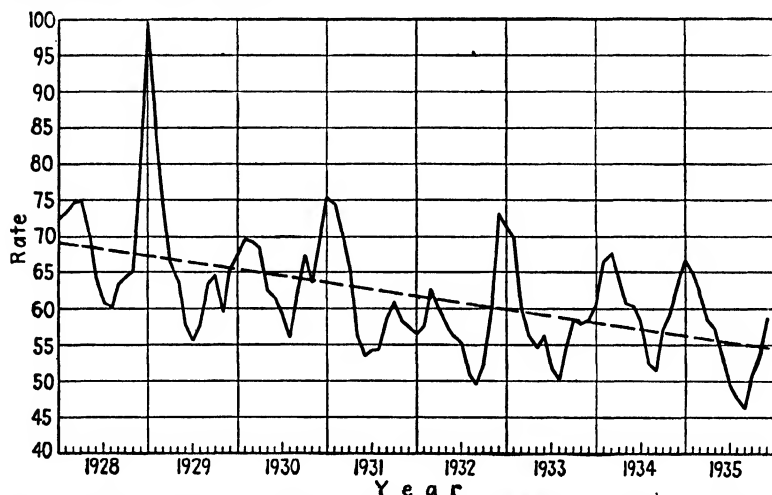
Fig. 58.—Monthly infant mortality rates in the registration area of the United States, 1928–1935. (From Table 84.)

Since in each year the infant death rate declines on the average 1.864, in one month the decline is 1.864/12 = 0.1553. In Table 85 we used average annual rates, which apply to the middle of a year. The middle of the year falls on June 30. The average monthly rates, however, apply to the middle of each month. We, therefore, enter Table 84 at June, 1928, and add to the annual 1928 trend rate of 68.388 one-half of the correction factor, 0.1553, so that we have

$$68.388 + .0777 = 68.4657$$

as the June, 1928, monthly trend in col. (3) of Table 84. We then add 0.1553 accumulatively to this rate for the five preceding months in 1928, and subtract 0.1553 accumulatively from it for each subsequent month throughout the eight-year period, which

completes col. (3). The monthly trend line, which is identical with the annual trend line, and the observed monthly rates from col. (2) of Table 84, are plotted in Fig. 58. From this graph, it is seen that in spite of a general downward trend in infant mortality rates, these rates have fluctuated considerably, so that even in, say, early 1935 they were much higher than in the middle of 1928. How much of this variation is due to the season of the year?

TABLE 86.—FREQUENCY DISTRIBUTION OF OBSERVED INFANT MORTALITY RATES EXPRESSED AS PERCENTAGES OF TREND, BY MONTHS, UNITED STATES REGISTRATION AREA, 1928–1935*

Observed rates, per cent of trend	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
145–149	/											
140–144												
135–139					.							
130–134												
125–129		/										
120–124												//
115–119	///	///	/									
110–114		/	///	/								/
105–109		//	//	//	/	/						//
100–104	///		//	//	//		/			///	//	/
95– 99				//	//	//			//	//	////	/
90– 94	/	/		/	//	///	//	/	////	//	//	/
85– 89				/	/	////	⊬⊬ /			/		
80– 84						/	/	/	//			

* From col. (4), Table 84.

Column (4) of Table 84 shows the monthly observed rates expressed as percentages of the monthly secular trend rates. These percentages represent the seasonal variations combined with the short-term cycles and the random fluctuations, but with the secular trend eliminated. To remove the short-term

cycles and random fluctuations, it is necessary to average the percentages for each month, over the eight-year period. As an aid in revealing whether or not a seasonal movement actually exists, and in choosing the most stable kind of monthly average, Table 86 is set up. A glance at it shows clearly the presence of a seasonal pattern in infant mortality. The death rate is high in the winter and low in the late summer. From the arrangement of the frequencies in the several columns, it appears that, except possibly in January, the arithmetic mean is a suitable average to use in this case. As a rule, however, it is recommended to
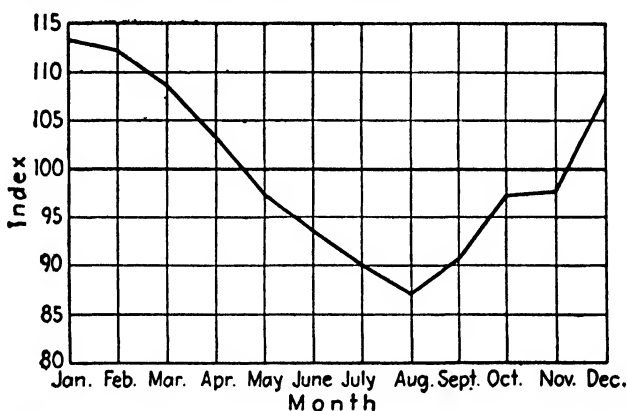
Fig. 59.—Seasonal indexes of infant mortality rates in the registration area of the United States, 1928–1935. (From Table 87.)

average the middle three or four values for each month, a sort of combined mean and median average which avoids the distortion due to extreme values. In Table 87 the mean monthly values are found[1] and are entered in col. (5) of Table 84. To convert the 12 mean monthly percentages to index numbers, they are divided by their own average, 99.99, and the quotient multiplied by 100, to give the last row of Table 87 and col. (6) of Table 84. The index of seasonal variation has an advantage over the simple percentages of col. (5) of Table 84, in that they vary around a mean of exactly 100.00 per cent, and are therefore more generally comparable and finished in form. In the seasonal indexes of col. (6) of Table 84 there now remains only the seasonal variation, since the secular trend, cycles, and random fluctuations were removed by the steps just taken. An undistorted

[1] From col. (4) of Table 84.

idea of the seasonal variation can now be obtained by plotting the monthly seasonal indexes around their mean of 100 per cent, as in Fig. 59. It is again obvious that the winter months are the danger period for infants.

6. Short-term Cycles Freed from Seasonal Fluctuations.—If it is wanted to observe the short-time cycles mixed with random fluctuations in the monthly infant mortality rates, freed from
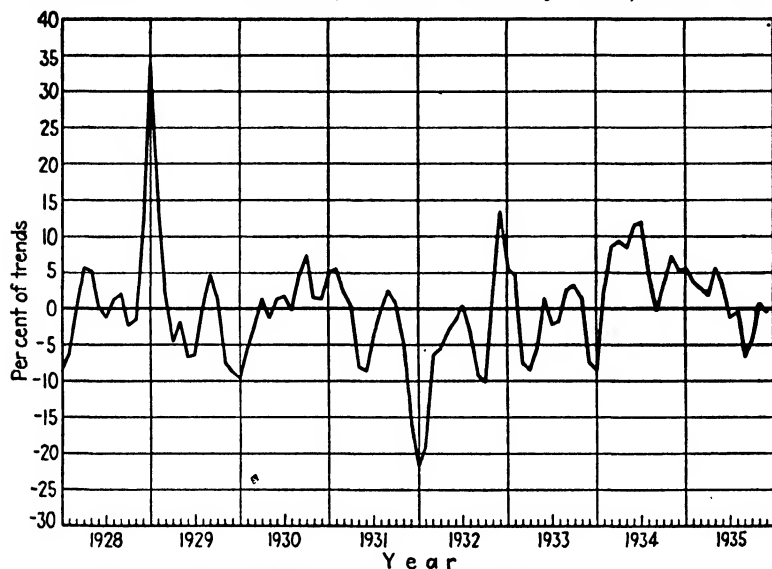
Fig. 60.—Short-term cycles and random fluctuations in infant mortality rates, United States registration area, 1928–1935. (From Table 84.)

both the secular trend and the seasonal movement, this may be done by recording in col. (7) of Table 84 the differences between the percentages of trend in col. (4) and the seasonal indexes in col. (6), and plotting them in Fig. 60. It appears that a number of other factors besides the season of the year affect the infant death rate, and need to be studied and brought under control. There is no suggestion from Fig. 60 that any progress was made during the eight-year period in reducing the percentage of infant deaths due to cyclical and random causes. The point might be tested by obtaining the standard deviations around zero of the differences in col. (7) of Table 84, for the first two years and the last two years of the period, and comparing the two standard deviations.

TABLE 87.—CALCULATION OF MONTHLY MEANS OF INFANT MORTALITY
RATES EXPRESSED AS PERCENTAGES OF TREND, UNITED STATES
REGISTRATION AREA, 1928–1935

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1928	104.56	105.95	108.52	109.04	102.59	93.76	89.01	88.32	93.24	94.78	96.32	120.39
1929	147.05	126.15	110.78	98.79	95.72	86.79	83.82	87.04	95.87	98.36	90.25	99.28
1930	103.50	106.79	106.29	104.84	96.32	94.84	91.82	86.92	96.00	104.65	99.28	109.39
1931	118.30	117.50	111.15	103.99	89.48	84.94	86.26	86.80	93.73	97.99	93.56	92.51
1932	91.44	93.30	102.15	97.85	94 49	91.95	90.71	83.86	82.09	86.93	100.28	121.50
1933	118.83	116.95	100.82	94.69	92.24	95.01	87.98	85.15	93.56	100.31	99.37	100.50
1934	104.37	114.85	117.23	112.69	105.85	105.10	102.05	91.63	90.81	100.60	104.94	113.22
1935	118.70	115.99	111.49	105.15	103.11	96.36	89.03	86.55	84.26	93.07	98.65	107.73
Total...........	906.75	897.48	868.43	827.04	779.80	748.75	720.68	696.27	729.56	776.69	782.65	864.52
Mean...........	113.34	112.19	108.55	103.38	97.48	93.59	90.09	87.03	91.20	97.09	97.83	108.07
Index...........	113.35	112.20	108.56	103.39	97.49	93.60	90.10	87.04	91.21	97.10	97.84	108.08

Exercises

1. Compare the trends in the birth rates of cities and of rural areas
in the original registration area of the United States over the 19-year
period, 1915 through 1933, using the data in the table below. Show
the cyclical deviations from trend, compare the variability of the two
series, and calculate the amount of correlation between the fluctuations
of the two series. What should be done with the data for extremely
atypical years, such as the war year, 1918? Is the correlation improved
by "lagging" one of the series? Plot all data.

BIRTH RATES PER 1,000 POPULATION FOR CITIES AND RURAL AREAS IN THE
ORIGINAL REGISTRATION AREA OF THE UNITED STATES, 1915–1933*

Year	Birth rate		Year	Birth rate	
	Cities	Rural		Cities	Rural
1915	26.0	23.8	1925	22.2	20.3
1916	26.0	23.5	1926	21.5	19.1
1917	26.4	23.3	1927	21.3	19.0
1918	25.8	23.0	1928	20.5	18.1
1919	23.8	21.1	1929	19.7	16.8
1920	24.6	22.2	1930	19.3	16.7
1921	24.5	23.1	1931	17.8	16.3
1922	22.9	21.8	1932	17.0	15.5
1923	22.9	21.1	1933	15.8	14.8
1924	23.2	21.2			

* From *Birth, Stillbirth, and Infant Mortality Statistics*, 1935, pp. 5–6, Bureau of the
Census.

2. For the relief data in the accompanying table show the secular trend and the seasonal fluctuations, and plot the results in each case.

NUMBER OF CASES RECEIVING RELIEF IN 385 RURAL AND TOWN AREAS OF THE UNITED STATES, 1932–1936*

Month	Cases				
	1932	1933	1934	1935	1936
January..............	30,931	99,064	169,554	298,785	145,734
February............	32,552	107,860	177,041	299,217	146,697
March..............	34,239	128,794	202,551	290,217	143,000
April...............	32,965	121,234	216,463	279,901	131,038
May................	30,713	112,079	222,647	266,014	123,102
June................	30,774	110,158	232,331	244,074	117,808
July................	29,687	131,850	239,441	227,814	120,067
August..............	30,214	126,572	259,410	218,883	128,303
September...........	33,561	114,147	255,929	204,745	129,124
October.............	38,126	117,459	251,397	201,341	144,492
November...........	65,922	135,234	262,635	198,780	149,781
December...........	75,517	115,877	282,068	167,297	166,173

* Adapted from Waller Wynne, Jr., *Five Years of Rural Relief*, p. 36, WPA, Division of Social Research, 1938.

References

CHADDOCK, R. E.: *Principles and Methods of Statistics*, Chap. XIII, Houghton Mifflin Company, Boston, 1925.

CROXTON, F. E., and D. J. COWDEN: *Applied General Statistics*, Chaps. XIV–XIX and XXV, Prentice-Hall, Inc., New York, 1939.

DAVIES, G. R., and DALE YODER: *Business Statistics*, Chaps. IV and V, John Wiley & Sons, Inc., New York, 1937.

MILLS, F. C.: *Statistical Methods*, rev. ed., Chaps. VII, VIII, and XI, Henry Holt and Company, Inc., New York, 1938.

WAUGH, A. E.: *Elements of Statistical Method*, Chap. VIII, McGraw-Hill Book Company, Inc., New York, 1938.

WHITE, R. C.: *Social Statistics*, Chap. XIII, Harper & Brothers, New York, 1933.

Appendix

TABLE 1.—AREA AND ORDINATE OF THE NORMAL CURVE[1]

x/σ	Area	Ordinate (y)	x/σ	Area	Ordinate (y)
.00	.0000000	.3989423	.46	.1772419	.3588903
.01	.0039894	.3989223	.47	.1808225	.3572253
.02	.0079783	.3988625	.48	.1843863	.3555325
.03	.0119665	.3987628	.49	.1879331	.3538124
.04	.0159534	.3986233	.50	.1914625	.3520653
.05	.0199388	.3984439			
			.51	.1949743	.3502919
.06	.0239222	.3982248	.52	.1984682	.3484925
.07	.0279032	.3979661	.53	.2019440	.3466677
.08	.0318814	.3976677	.54	.2054015	.3448180
.09	.0358564	.3973298	.55	.2088403	.3429439
.10	.0398278	.3969525			
			.56	.2122603	.3410458
.11	.0437953	.3965360	.57	.2156612	.3391243
.12	.0477584	.3960802	.58	.2190427	.3371799
.13	.0517168	.3955854	.59	.2224047	.3352132
.14	.0556700	.3950517	.60	.2257469	.3332246
.15	.0596177	.3944793			
			.61	.2290691	.3312147
.16	.0635595	.3938684	.62	.2323711	.3291840
.17	.0674949	.3932190	.63	.2356527	.3271330
.18	.0714237	.3925315	.64	.2389137	.3250623
.19	.0753454	.3918060	.65	.2421539	.3229724
.20	.0792597	.3910427			
			.66	.2453731	.3208638
.21	.0831662	.3902419	.67	.2485711	.3187371
.22	.0870644	.3894038	.68	.2517478	.3165929
.23	.0909541	.3885286	.69	.2549029	.3144317
.24	.0948349	.3876166	.70	.2580363	.3122539
.25	.0987063	.3866681			
			.71	.2611479	.3100603
.26	.1025681	.3856834	.72	.2642375	.3078513
.27	.1064199	.3846627	.73	.2673049	.3056274
.28	.1102612	.3836063	.74	.2703500	.3033893
.29	.1140919	.3825146	.75	.2733726	.3011374
.30	.1179114	.3813878			
			.76	.2763727	.2988724
.31	.1217195	.3802264	.77	.2793501	.2965948
.32	.1255158	.3790305	.78	.2823046	.2943050
.33	.1293000	.3778007	.79	.2852361	.2920038
.34	.1330717	.3765372	.80	.2881446	.2896916
.35	.1368307	.3752403			
			.81	.2910299	.2873689
.36	.1405764	.3739106	.82	.2938919	.2850364
.37	.1443088	.3725483	.83	.2967306	.2826945
.38	.1480273	.3711539	.84	.2995458	.2803438
.39	.1517317	.3697277	.85	.3023375	.2779849
.40	.1555417	.3682707			
			.86	.3051055	.2756182
.41	.1590970	.3667817	.87	.3078498	.2732444
.42	.1627573	.3652627	.88	.3105703	.2708640
.43	.1664022	.3637136	.89	.3132671	.2684774
.44	.1700314	.3621349	.90	.3159399	.2660852
.45	.1736448	.3605270			

[1] From Kent, "The Elements of Statistics."

TABLE 1.—AREA AND ORDINATE OF THE NORMAL CURVE.[1]—(*Continued*)

x/σ	Area	Ordinate (y)	x/σ	Area	Ordinate (y)
.91	.3185887	.2636880	1.36	.4130850	.1582248
.92	.3212136	.2612863	1.37	.4146565	.1560797
.93	.3238145	.2588805	1.38	.4162067	.1539483
.94	.3263912	.2564713	1.39	.4177356	.1518308
.95	.3289439	.2540591	1.40	.4192433	.1497275
.96	.3314724	.2516443	1.41	.4207302	.1476385
.97	.3339768	.2492277	1.42	.4221962	.1455641
.98	.3364569	.2468095	1.43	.4236415	.1435046
.99	.3389129	.2443904	1.44	.4250663	.1414600
1.00	.3413447	.2419707	1.45	.4264707	.1394306
1.01	.3437524	.2395511	1.46	.4278550	.1374165
1.02	.3461358	.2371320	1.47	.4292191	.1354181
1.03	.3484950	.2347138	1.48	.4305634	.1334353
1.04	.3508300	.2322970	1.49	.4318879	.1314684
1.05	.3531409	.2298821	1.50	.4331928	.1295176
1.06	.3554277	.2274696	1.51	.4344783	.1275830
1.07	.3576903	.2250599	1.52	·.4357445	.1256646
1.08	.3599289	.2226535	1.53	.4369916	.1237628
1.09	.3621434	.2202508	1.54	.4382198	.1218775
1.10	.3643339	.2178522	1.55	.4394292	.1200090
1.11	.3665005	.2154582	1.56	.4406201	.1181573
1.12	.3686431	.2130691	1.57	.4417924	.1163225
1.13	.3707619	.2106856	1.58	.4429466	.1145048
1.14	.3728568	.2083078	1.59	.4440826	.1127042
1.15	.3749281	.2059363	1.60	.4452007	.1109208
1.16	.3769756	.2035714	1.61	.4463011	.1091548
1.17	.3789995	.2012135	1.62	.4473839	.1074061
1.18	.3809999	.1988631	1.63	.4484493	.1056748
1.19	.3829768	.1965205	1.64	.4494974	.1039611
1.20	.3849303	.1941861	1.65	.4505285	.1022649
1.21	.3868606	.1918602	1.66	.4515428	.1005864
1.22	.3887676	.1895432	1.67	.4525403	.0989255
1.23	.3906514	.1872354	1.68	.4535213	.0972823
1.24	.3925123	.1849373	1.69	.4544860	.0956568
1.25	.3943502	.1826491	1.70	.4554345	.0940491
1.26	.3961653	.1803712	1.71	.4563671	.0924591
1.27	.3979577	.1781038	1.72	.4572838	.0908870
1.28	.3997274	.1758474	1.73	.4581849	.0893326
1.29	.4014747	.1736022	1.74	.4590705	.0877961
1.30	.4031995	.1713686	1.75	.4599408	.0862773
1.31	.4049021	.1691468	1.76	.4607961	.0847764
1.32	.4065825	.1669370	1.77	.4616364	.0832932
1.33	.4082409	.1647397	1.78	.4624620	.0818278
1.34	.4098773	.1625551	1.79	.4632730	.0803801
1.35	.4114920	.1603833	1.80	.4640697	.0789502

[1] From Kent, "The Elements of Statistics."

TABLE 1.—AREA AND ORDINATE OF THE NORMAL CURVE.[1]—*(Continued)*

x/σ	Area	Ordinate (y)	x/σ	Area	Ordinate (y)
1.81	.4648521	.0775379	2.26	.4880894	.0310319
1.82	.4656205	.0761433	2.27	.4883962	.0303370
1.83	.4663750	.0747663	2.28	.4886962	.0296546
1.84	.4671159	.0734068	2.29	.4889893	.0289847
1.85	.4678432	.0720649	2.30	.4892759	.0283270
1.86	.4685572	.0707404	2.31	.4895559	.0276816
1.87	.4692581	.0694333	2.32	.4898296	.0270481
1.88	.4699460	.0681436	2.33	.4900969	.0264265
1.89	.4706210	.0668711	2.34	.4903581	.0258166
1.90	.4712834	.0656158	2.35	.4906133	.0252182
1.91	.4719334	.0643777	2.36	.4908625	.0246313
1.92	.4725711	.0631566	2.37	.4911060	.0240556
1.93	.4731966	.0619524	2.38	.4913437	.0234910
1.94	.4738102	.0607652	2.39	.4915758	.0229374
1.95	.4744119	.0595947	2.40	.4918025	.0223945
1.96	.4750021	.0584409	2.41	.4920237	.0218624
1.97	.4755808	.0573038	2.42	.4922397	.0213407
1.98	.4761482	.0561831	2.43	.4924506	.0208294
1.99	.4767045	.0550789	2.44	.4926564	.0203284
2.00	.4772499	.0539910	2.45	.4928572	.0198374
2.01	.4777844	.0529192	2.46	.4930531	.0193563
2.02	.4783083	.0518636	2.47	.4932443	.0188850
2.03	.4788217	.0508239	2.48	.4934309	.0184233
2.04	.4793248	.0498001	2.49	.4936128	.0179711
2.05	.4798178	.0487920	2.50	.4937903	.0175283
2.06	.4803007	.0477996	2.51	.4939634	.0170947
2.07	.4807738	.0468226	2.52	.4941323	.0166701
2.08	.4812372	.0458611	2.53	.4943001	.0162452
2.09	.4816911	.0449148	2.54	.4944574	.0158476
2.10	.4821356	.0439836	2.55	.4946139	.0154493
2.11	.4825708	.0430674	2.56	.4947664	.0150596
2.12	.4829970	.0421661	2.57	.4949151	.0146782
2.13	.4834142	.0412795	2.58	.4950600	.0143051
2.14	.4838226	.0404076	2.59	.4952012	.0139401
2.15	.4842224	.0395500	2.60	.4953388	.0135830
2.16	.4846137	.0387069	2.61	.4954729	.0132337
2.17	.4849966	.0378779	2.62	.4956035	.0128921
2.18	.4853713	.0370629	2.63	.4957308	.0125581
2.19	.4857379	.0362619	2.64	.4958547	.0122315
2.20	.4860966	.0354746	2.65	.4959754	.0119122
2.21	.4864474	.0347009	2.66	.4960930	.0116001
2.22	.4867906	.0339408	2.67	.4962074	.0112951
2.23	.4871263	.0331939	2.68	.4963189	.0109969
2.24	.4874545	.0324603	2.69	.4964274	.0107056
2.25	.4877755	.0317397	2.70	.4965330	.0104209

[1] From Kent, "The Elements of Statistics."

TABLE 1.—AREA AND ORDINATE OF THE NORMAL CURVE.[1]—(*Continued*)

x/σ	Area	Ordinate (y)	x/σ	Area	Ordinate (y)
2.71	.4966358	.0101428	3.16	.4992112	.0027075
2.72	.4967359	.0098712	3.17	.4992378	.0026231
2.73	.4968333	.0096058	3.18	.4992636	.0025412
2.74	.4969280	.0093466	3.19	.4992886	.0024615
2.75	.4970202	.0090936	3.20	.4993129	.0023841
2.76	.4971099	.0088465	3.21	.4993363	.0023089
2.77	.4971972	.0086052	3.22	.4993593	.0022358
2.78	.4972821	.0083697	3.23	.4993810	.0021649
2.79	.4973646	.0081398	3.24	.4994024	.0020960
2.80	.4974449	.0079155	3.25	.4994230	.0020290
2.81	.4975229	.0076965	3.26	.4994429	.0019641
2.82	.4975988	.0074829	3.27	.4994623	.0019010
2.83	.4976726	.0072744	3.28	.4994810	.0018397
2.84	.4977443	.0070711	3.29	.4994991	.0017803
2.85	.4978140	.0068728	3.30	.4995166	.0017226
2.86	.4978818	.0066793	3.31	.4995335	.0016666
2.87	.4979476	.0064907	3.32	.4995658	.0016122
2.88	.4980116	.0063067	3.33	.4995658	.0015595
2.89	.4980738	.0061274	3.34	.4995811	.0015084
2.90	.4981342	.0059525	3.35	.4995959	.0014587
2.91	.4981929	.0057821	3.36	.4996103	.0014106
2.92	.4982498	.0056160	3.37	.4996242	.0013639
2.93	.4983052	.0054541	3.38	.4996376	.0013187
2.94	.4983589	.0052963	3.39	.4996505	.0012748
2.95	.4984111	.0051426	3.40	.4996631	.0012322
2.96	.4984618	.0049929	3.41	.4996752	.0011910
2.97	.4985110	.0048470	3.42	.4996869	.0011510
2.98	.4985588	.0047050	3.43	.4996982	.0011122
2.99	.4986051	.0045666	3.44	.4997091	.0010747
3.00	.4986501	.0044318	3.45	.4997197	.0010383
3.01	.4986938	.0043007	3.46	.4997299	.0010030
3.02	.4987361	.0041729	3.47	.4997398	.0009689
3.03	.4987772	.0040486	3.48	.4997493	.0009358
3.04	.4988171	.0039276	3.49	.4997585	.0009037
3.05	.4988558	.0038098	3.50	.4997674	.0008727
3.06	.4988933	.0036951	3.51	.4997759	.0008426
3.07	.4989297	.0035836	3.52	.4997842	.0008135
3.08	.4989650	.0034751	3.53	.4997922	.0007853
3.09	.4989992	.0033695	3.54	.4997999	.0007581
3.10	.4990324	.0032668	3.55	.4998074	.0007317
3.11	.4990646	.0031669	3.56	.4998146	.0007061
3.12	.4990957	.0030698	3.57	.4998215	.0006814
3.13	.4991260	.0029754	3.58	.4998282	.0006575
3.14	.4991553	.0028835	3.59	.4998347	.0006343
3.15	.4991836	.0027943	3.60	.4998409	.0006119

[1] From Kent, "The Elements of Statistics."

TABLE 1.—AREA AND ORDINATE OF THE NORMAL CURVE.[1]—(*Continued*)

x/σ	Area	Ordinate (y)	x/σ	Area	Ordinate (y)
3.61	.4998469	.0005902	4.06	.4999755	.0001051
3.62	.4998527	.0005693	4.07	.4999765	.0001009
3.63	.4998583	.0005490	4.08	.4999775	.0000969
3.64	.4998637	.0005294	4.09	.4999784	.0000930
3.65	.4998689	.0005105	4.10	.4999793	.0000893
3.66	.4998739	.0004921	4.11	.4999802	.0000857
3.67	.4998787	.0004744	4.12	.4999811	.0000822
3.68	.4998834	.0004573	4.13	.4999819	.0000789
3.69	.4998879	.0004408	4.14	.4999826	.0000757
3.70	.4998922	.0004248	4.15	.4999834	.0000726
3.71	.4998964	.0004093	4.16	.4999841	.0000697
3.72	.4999004	.0003800	4.17	.4999848	.0000668
3.73	.4999043	.0003661	4.18	.4999854	.0000641
3.74	.4999080	.0003526	4.19	.4999861	.0000615
3.75	.4999116	.0003386	4.20	.4999867	.0000589
3.76	.4999150	.0003396	4.21	.4999872	.0000565
3.77	.4999184	.0003271	4.22	.4999878	.0000542
3.78	.4999216	.0003149	4.23	.4999883	.0000519
3.79	.4999247	.0003032	4.24	.4999888	.0000498
3.80	.4999277	.0002919	4.25	.4999893	.0000477
3.81	.4999305	.0002810	4.26	.4999898	.0000457
3.82	.4999333	.0002705	4.27	.4999902	.0000438
3.83	.4999359	.0002604	4.28	.4999907	.0000420
3.84	.4999385	.0002506	4.29	.4999911	.0000402
3.85	.4999409	.0002411	4.30	.4999915	.0000385
3.86	.4999433	.0002320	4.31	.4999918	.0000369
3.87	.4999456	.0002232	4.32	.4999922	.0000354
3.88	.4999478	.0002147	4.33	.4999925	.0000339
3.89	.4999499	.0002065	4.34	.4999929	.0000324
3.90	.4999519	.0001987	4.35	.4999932	.0000310
3.91	.4999539	.0001910	4.36	.4999935	.0000297
3.92	.4999557	.0001837	4.37	.4999938	.0000284
3.93	.4999575	.0001766	4.38	.4999941	.0000272
3.94	.4999593	.0001698	4.39	.4999943	.0000261
3.95	.4999609	.0001633	4.40	.4999946	.0000249
3.96	.4999625	.0001569	4.41	.4999948	.0000239
3.97	.4999641	.0001508	4.42	.4999951	.0000228
3.98	.4999655	.0001449	4.43	.4999953	.0000218
3.99	.4999670	.0001393	4.44	.4999955	.0000209
4.00	.4999683	.0001338	4.45	.4999957	.0000200
4.01 *c*	.4999696	.0001286	4.46	.4999959	.0000191
4.02	.4999709	.0001235	4.47	.4999961	.0000183
4.03	.4999721	.0001186	4.48	.4999963	.0000175
4.04	.4999733	.0001140	4.49	.4999964	.0000167
4.05	.4999744	.0001094	4.50	.4999966	.0000160

[1] From Kent, "The Elements of Statistics."

TABLE 2.—TABLE OF χ² (CHI-SQUARE)*

n	P = .99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n - 1}$ may be used as a normal deviate with unit standard deviation.

* This table is taken by consent from *Statistical Methods for Research Workers* by Prof. R. A. Fisher, by Oliver & Boyd, Edinburgh, and attention is drawn to the larger collection in *Statistical Tables* by Prof. R. A. Fisher and F. Yates, by Oliver & Boyd, Edinburgh.

TABLE 3.—BINOMIAL COEFFICIENTS, $_nC_r$

n \ r	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1														
2	1	2	1													
3	1	3	3	1												
4	1	4	6	4	1											
5	1	5	10	10	5	1										
6	1	6	15	20	15	6	1									
7	1	7	21	35	35	21	7	1								
8	1	8	28	56	70	56	28	8	1							
9	1	9	36	84	126	126	84	36	9	1						
10	1	10	45	120	210	252	210	120	45	10	1					
11	1	11	55	165	330	462	462	330	165	55	11	1				
12	1	12	66	220	495	792	924	792	495	220	66	12	1			
13	1	13	78	286	715	1287	1716	1716	1287	715	286	78	13	1		
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001	364	91	14	1	
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003	1365	455	105	15	1

TABLE 4.—VALUES OF THE CORRELATION COEFFICIENT FOR DIFFERENT
LEVELS OF SIGNIFICANCE*

P n	.05	.01
1	.996917	.9998766
2	.95000	.990000
3	.8783	.95873
4	.8114	.91720
5	.7545	.8745
6	.7067	.8343
7	.6664	.7977
8	.6319	.7646
9	.6021	.7348
10	.5760	.7079
11	.5529	.6835
12	.5324	.6614
13	.5139	.6411
14	.4973	.6226
15	.4821	.6055
16	.4683	.5897
17	.4555	.5751
18	.4438	.5614
19	.4329	.5487
20	.4227	.5368
25	.3809	.4869
30	.3494	.4487
35	.3246	.4182
40	.3044	.3932
45	.2875	.3721
50	.2732	.3541
60	.2500	.3248
70	.2319	.3017
80	.2172	.2830
90	.2050	.2673
100	.1946	.2540

For a total correlation, n is 2 less than the number of pairs in the sample; for a partial correlation, the number of eliminated variates also should be subtracted.

* This table is taken by consent from *Statistical Methods for Research Workers* by Prof. R. A. Fisher, by Oliver & Boyd, Edinburgh, and attention is drawn to the larger collection in *Statistical Tables* by Prof. R. A. Fisher and F. Yates, by Oliver & Boyd, Edinburgh

TABLE 5.—VALUES OF z FOR GIVEN VALUES OF r*

r	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.000	.0000	.0010	.0020	.0030	.0040	.0050	.0060	.0070	.0080	.0090
.010	.0100	.0110	.0120	.0130	.0140	.0150	.0160	.0170	.0180	.0190
.020	.0200	.0210	.0220	.0230	.0240	.0250	.0260	.0270	.0280	.0290
.030	.0300	.0310	.0320	.0330	.0340	.0350	.0360	.0370	.0380	.0390
.040	.0400	.0410	.0420	.0430	.0440	.0450	.0460	.0470	.0480	.0490
.050	.0501	.0511	.0521	.0531	.0541	.0551	.0561	.0571	.0581	.0591
.060	.0601	.0611	.0621	.0631	.0641	.0651	.0661	.0671	.0681	.0691
.070	.0701	.0711	.0721	.0731	.0741	.0751	.0761	.0771	.0782	.0792
.080	.0802	.0812	.0822	.0832	.0842	.0852	.0862	.0872	.0882	.0892
.090	.0902	.0912	.0922	.0933	.0943	.0953	.0963	.0973	.0983	.0993
.100	.1003	.1013	.1024	.1034	.1044	.1054	.1064	.1074	.1084	.1094
.110	.1105	.1115	.1125	.1135	.1145	.1155	.1165	.1175	.1185	.1195
.120	.1206	.1216	.1226	.1236	.1246	.1257	.1267	.1277	.1287	.1297
.130	.1308	.1318	.1328	.1338	.1348	.1358	.1368	.1379	.1389	.1399
.140	.1409	.1419	.1430	.1440	.1450	.1460	.1470	.1481	.1491	.1501
.150	.1511	.1522	.1532	.1542	.1552	.1563	.1573	.1583	.1593	.1604
.160	.1614	.1624	.1634	.1654	.1655	.1665	.1676	.1686	.1696	.1706
.170	.1717	.1727	.1737	.1748	.1758	.1768	.1779	.1789	.1799	.1810
.180	.1820	.1830	.1841	.1851	.1861	.1872	.1882	.1892	.1903	.1913
.190	.1923	.1934	.1944	.1954	.1965	.1975	.1986	.1996	.2007	.2017
.200	.2027	.2038	.2048	.2059	.2069	.2079	.2090	.2100	.2111	.2121
.210	.2132	.2142	.2153	.2163	.2174	.2184	.2194	.2205	.2215	.2226
.220	.2237	.2247	.2258	.2268	.2279	.2289	.2300	.2310	.2321	.2331
.230	.2342	.2353	.2363	.2374	.2384	.2395	.2405	.2416	.2427	.2437
.240	.2448	.2458	.2469	.2480	.2490	.2501	.2511	.2522	.2533	.2543
.250	.2554	.2565	.2575	.2586	.2597	.2608	.2618	.2629	.2640	.2650
.260	.2661	.2672	.2682	.2693	.2704	.2715	.2726	.2736	.2747	.2758
.370	.2769	.2779	.2790	.2801	.2812	.2823	.2833	.2844	.2855	.2866
.280	.2877	.2888	.2898	.2909	.2920	.2931	.2942	.2953	.2964	.2975
.290	.2986	.2997	.3008	.3019	.3029	.3040	.3051	.3062	.3073	.3084
.300	.3095	.3106	.3117	.3128	.3139	.3150	.3161	.3172	.3183	.3195
.310	.3206	.3217	.3228	.3239	.3250	.3261	.3272	.3283	.3294	.3305
.320	.3317	.3328	.3339	.3350	.3361	.3372	.3384	.3395	.3406	.3417
.330	.3428	.3439	.3451	.3462	.3473	.3484	.3496	.3507	.3518	.3530
.340	.3541	.3552	.3564	.3575	.3586	.3597	.3609	.3620	.3632	.3643
.350	.3654	.3666	.3677	.3689	.3700	.3712	.3723	.3734	.3746	.3757
.360	.3769	.3780	.3792	.3803	.3815	.3826	.3838	.3850	.3861	.3873
.370	.3884	.3896	.3907	.3919	.3931	.3942	.3954	.3966	.3977	.3989
.380	.4001	.4012	.4024	.4036	.4047	.4059	.4071	.4083	.4094	.4106
.390	.4118	.4130	.4142	.4153	.4165	.4177	.4189	.4201	.4213	.4225
.400	.4236	.4248	.4260	.4272	.4284	.4296	.4308	.4320	.4332	.4344
.410	.4356	.4368	.4380	.4392	.4404	.4416	.4429	.4441	.4453	.4465
.420	.4477	.4489	.4501	.4513	.4526	.4538	.4550	.4562	.4574	.4587
.430	.4599	.4611	.4623	.4636	.4648	.4660	.4673	.4685	.4697	.4710
.440	.4722	.4735	.4747	.4760	.4772	.4784	.4797	.4809	.4822	.4835
.450	.4847	.4860	.4872	.4885	.4897	.4910	.4923	.4935	.4948	.4961
.460	.4973	.4986	.4999	.5011	.5024	.5037	.5049	.5062	.5075	.5088
.470	.5101	.5114	.5126	.5139	.5152	.5165	.5178	.5191	.5204	.5217
.480	.5230	.5243	.5256	.5279	.5282	.5295	.5308	.5321	.5334	.5347
.490	.5361	.5374	.5387	.5400	.5413	.5427	.5440	.5453	.5466	.5480

* From Albert E. Waugh, *Laboratory Manual and Problems for Elements of Statistical Method*, pp. 32–33, McGraw-Hill Book Company, Inc., New York.

TABLE 5.—VALUES OF z FOR GIVEN VALUES OF r.—(Continued)

r	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.500	.5493	.5506	.5520	.5533	.5547	.5560	.5573	.5587	.5600	.5614
.510	.5627	.5641	.5654	.5668	.5681	.5695	.5709	.5722	.5736	.5750
.520	.5763	.5777	.5791	.5805	.5818	.5832	.5846	.5860	.5874	.5888
.530	.5901	.5915	.5929	.5943	.5957	.5971	.5985	.5999	.6013	.6027
.540	.6042	.6056	.6070	.6084	.6098	.6112	.6127	.6141	.6155	.6170
.550	.6184	.6198	.6213	.6227	.6241	.6256	.6270	.6285	.6299	.6314
.560	.6328	.6343	.6358	.6372	.6387	.6401	.6416	.6431	.6446	.6460
.570	.6475	.6490	.6505	.6520	.6535	.6550	.6565	.6579	.6594	.6610
.580	.6625	.6640	.6655	.6670	.6685	.6700	.6715	.6731	.6746	.6761
.590	.6777	.6792	.6807	.6823	.6838	.6854	.6869	.6885	.6900	.6916
.600	.6931	.6947	.6963	.6978	.6994	.7010	.7026	.7042	.7057	.7073
.610	.7089	.7105	.7121	.7137	.7153	.7169	.7185	.7201	.7218	.7234
.620	.7250	.7266	.7283	.7299	.7315	.7332	.7348	.7364	.7381	.7398
.630	.7414	.7431	.7447	.7464	.7481	.7497	.7514	.7531	.7548	.7565
.640	.7582	.7599	.7616	.7633	.7650	.7667	.7684	.7701	.7718	.7736
.650	.7753	.7770	.7788	.7805	.7823	.7840	.7858	.7875	.7893	.7910
.660	.7928	.7946	.7964	.7981	.7999	.8017	.8035	.8053	.8071	.8089
.670	.8107	.8126	.8144	.8162	.8180	.8199	.8217	.8236	.8254	.8273
.680	.8291	.8310	.8328	.8347	.8366	.8385	.8404	.8423	.8442	.8461
.690	.8480	.8499	.8518	.8537	.8556	.8576	.8595	.8614	.8634	.8653
.700	.8673	.8693	.8712	.8732	.8752	.8772	.8792	.8812	.8832	.8852
.710	.8872	.8892	.8912	.8933	.8953	.8973	.8994	.9014	.9035	.9056
.720	.9076	.9097	.9118	.9139	.9160	.9181	.9202	.9223	.9245	.9266
.730	.9287	.9309	.9330	.9352	.9373	.9395	.9417	.9439	.9461	.9483
.740	.9505	.9527	.9549	.9571	.9594	.9616	.9639	.9661	.9684	.9707
.750	.9730	.9752	.9775	.9799	.9822	.9845	.9868	.9892	.9915	.9939
.760	.9962	.9986	1.0010	1.0034	1.0058	1.0082	1.0106	1.0130	1.0154	1.0179
.770	1.0203	1.0228	1.0253	1.0277	1.0302	1.0327	1.0352	1.0378	1.0403	1.0428
.780	1.0454	1.0479	1.0505	1.0531	1.0557	1.0583	1.0609	1.0635	1.0661	1.0688
.790	1.0714	1.0741	1.0768	1.0795	1.0822	1.0849	1.0876	1.0903	1.0931	1.0958
.800	1.0986	1.1014	1.1041	1.1070	1.1098	1.1127	1.1155	1.1184	1.1212	1.1241
.810	1.1270	1.1299	1.1329	1.1358	1.1388	1.1417	1.1447	1.1477	1.1507	1.1538
.820	1.1568	1.1599	1.1630	1.1660	1.1692	1.1723	1.1754	1.1786	1.1817	1.1849
.830	1.1870	1.1913	1.1946	1.1979	1.2011	1.2044	1.2077	1.2111	1.2144	1.2178
.840	1.2212	1.2246	1.2280	1.2315	1.2349	1.2384	1.2419	1.2454	1.2490	1.2526
.850	1.2561	1.2598	1.2634	1.2670	1.2708	1.2744	1.2782	1.2819	1.2857	1.2895
.860	1.2934	1.2972	1.3011	1.3050	1.3089	1.3129	1.3168	1.3209	1.3249	1.3290
.870	1.3331	1.3372	1.3414	1.3456	1.3498	1.3540	1.3583	1.3626	1.3670	1.3714
.880	1.3758	1.3802	1.3847	1.3892	1.3938	1.3984	1.4030	1.4077	1.4124	1.4171
.890	1.4219	1.4268	1.4316	1.4366	1.4415	1.4465	1.4516	1.4566	1.4618	1.4670
.900	1.4722	1.4775	1.4828	1.4883	1.4937	1.4992	1.5047	1.5103	1.5160	1.5217
.910	1.5275	1.5334	1.5393	1.5453	1.5513	1.5574	1.5636	1.5698	1.5762	1.5825
.920	1.5890	1.5956	1.6022	1.6089	1.6157	1.6226	1.6296	1.6366	1.6438	1.6510
.930	1.6584	1.6659	1.6734	1.6811	1.6888	1.6967	1.7047	1.7129	1.7211	1.7295
.940	1.7380	1.7467	1.7555	1.7645	1.7736	1.7828	1.7923	1.8019	1.8117	1.8216
.950	1.8318	1.8421	1.8527	1.8635	1.8745	1.8857	1.8972	1.9090	1.9210	1.9333
.960	1.9459	1.9588	1.9721	1.9857	1.9996	2.0140	2.0287	2.0439	2.0595	2.0756
.970	2.0923	2.1095	2.1273	2.1457	2.1649	2.1847	2.2054	2.2269	2.2494	2.2729
.980	2.2976	2.3223	2.3507	2.3796	2.4101	2.4426	2.4774	2.5147	2.5550	2.5988
.990	2.6467	2.6996	2.7587	2.8257	2.9031	2.9945	3.1063	3.2504	3.4534	3.8002

r	z
.9999	4.95172
.99999	6.10303

Foreword to Table 6.—To extract the square root of any number, we begin at the decimal point and group the figures by pairs in both directions. For example, 7,500,000,000,000 becomes 07 50 00 00 00 00 00. In Table 6 we look up the figure, 750. Its square root is seen to be 27.3861. We allow one figure in the root for each pair of figures in the number. There are seven pairs to the left of the decimal in our number and none to the right of the decimal, so the root will contain seven figures to the left of the decimal, thus: 2,738,610. In looking up a square root, never separate the figures in a pair. In our illustration it would be wrong to find the square root of the number 75 or of the number 7500.

When the square root of a large number (*e.g.*, 7,583,615,000,000) cannot be found exactly from the table, the nearest approximation is often taken (*e.g.*, take the square root of 7,580,000,000,000 as roughly equivalent to the square root of 7,583,615,000,000). Where greater accuracy is required, a larger table may be used (see *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots, Reciprocals of All Integer Numbers Up to* 10,000, Spon and Chamberlain, 120 Liberty Street, New York), or any elementary textbook in algebra may be consulted for the method of extracting a square root. Calculating machine companies furnish pamphlets describing how to extract a square root on their machines. A slide rule gives approximate square roots easily and rapidly.

TABLE 6.—SQUARES AND SQUARE ROOTS*

Number	Square	Square root	Number	Square	Square root
1	1	1.0000	41	1681	6.4031
2	4	1.4142	42	1764	6.4807
3	9	1.7321	43	1849	6.5574
4	16	2.0000	44	1936	6.6332
5	25	2.2361	45	2025	6.7082
6	36	2.4495	46	2116	6.7823
7	49	2.6458	47	2209	6.8557
8	64	2.8284	48	2304	6.9282
9	81	3.0000	49	2401	7.0000
10	100	3.1623	50	2500	7.0711
11	121	3.3166	51	2601	7.1414
12	144	3.4641	52	2704	7.2111
13	169	3.6056	53	2809	7.2801
14	196	3.7417	54	2916	7.3485
15	225	3.8730	55	3025	7.4162
16	256	4.0000	56	3136	7.4833
17	289	4.1231	57	3249	7.5498
18	324	4.2426	58	3364	7.6158
19	361	4.3589	59	3481	7.6811
20	400	4.4721	60	3600	7.7460
21	441	4.5826	61	3721	7.8102
22	484	4.6904	62	3844	7.8740
23	529	4.7958	63	3969	7.9373
24	576	4.8990	64	4096	8.0000
25	625	5.0000	65	4225	8.0623
26	676	5.0990	66	4356	8.1240
27	729	5.1962	67	4489	8.1854
28	784	5.2915	68	4624	8.2462
29	841	5.3852	69	4761	8.3066
30	900	5.4772	70	4900	8.3666
31	961	5.5678	71	5041	8.4261
32	1024	5.6569	72	5184	8.4853
33	1089	5.7446	73	5329	8.5440
34	1156	5.8310	74	5476	8.6023
35	1225	5.9161	75	5625	8.6603
36	1296	6.0000	76	5776	8.7178
37	1369	6.0828	77	5929	8.7750
38	1444	6.1644	78	6084	8.8318
39	1521	6.2450	79	6241	8.8882
40	1600	6.3246	80	6400	8.9443

* From Herbert Sorenson, *Statistics for Students of Psychology and Education*, pp. 347–359, McGraw-Hill Book Company, Inc., New York

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
81	6561	9.0000	121	14641	11.0000
82	6724	9.0554	122	14884	11.0454
83	6889	9.1104	123	15129	11.0905
84	7056	9.1652	124	15376	11.1355
85	7225	9.2195	125	15625	11.1803
86	7396	9.2736	126	15876	11.2250
87	7569	9.3274	127	16129	11.2694
88	7744	9.3808	128	16384	11.3137
89	7921	9.4340	129	16641	11.3578
90	8100	9.4868	130	16900	11.4018
91	8281	9.5394	131	17161	11.4455
92	8464	9.5917	132	17424	11.4891
93	8649	9.6437	133	17689	11.5326
94	8836	9.6954	134	17956	11.5758
95	9025	9.7468	135	18225	11.6190
96	9216	9.7980	136	18496	11.6619
97	9409	9.8489	137	18769	11.7047
98	9604	9.8995	138	19044	11.7473
99	9801	9.9499	139	19321	11.7898
100	10000	10.0000	140	19600	11.8322
101	10201	10.0499	141	19881	11.8743
102	10404	10.0995	142	20164	11.9164
103	10609	10.1489	143	20449	11.9583
104	10816	10.1980	144	20736	12.0000
105	11025	10.2470	145	21025	12.0416
106	11236	10.2956	146	21316	12.0830
107	11449	10.3441	147	21609	12.1244
108	11664	10.3923	148	21904	12.1655
109	11881	10.4403	149	22201	12.2066
110	12100	10.4881	150	22500	12.2474
111	12321	10.5357	151	22801	12.2882
112	12544	10.5830	152	23104	12.3288
113	12769	10.6301	153	23409	12.3693
114	12996	10.6771	154	23716	12.4097
115	13225	10.7238	155	24025	12.4499
116	13456	10.7703	156	24336	12.4900
117	13689	10.8167	157	24649	12.5300
118	13924	10.8628	158	24964	12.5698
119	14161	10.9087	159	25281	12.6095
120	14400	10.9545	160	25600	12.6491

TABLE 6.—SQUARES AND SQUARE ROOTS.—*(Continued)*

Number	Square	Square root	Number	Square	Square root
161	25921	12.6886	201	40401	14.1774
162	26244	12.7279	202	40804	14.2127
163	26569	12.7671	203	41209	14.2478
164	26896	12.8062	204	41616	14.2829
165	27225	12.8452	205	42025	14.3178
166	27556	12.8841	206	42436	14.3527
167	27889	12.9228	207	42849	14.3875
168	28224	12.9615	208	43264	14.4222
169	28561	13.0000	209	43681	14.4568
170	28900	13.0384	210	44100	14.4914
171	29241	13.0767	211	44521	14.5258
172	29584	13.1149	212	44944	14.5602
173	29929	13.1529	213	45369	14.5945
174	30276	13.1909	214	45796	14.6287
175	30625	13.2288	215	46225	14.6629
176	30976	13.2665	216	46656	14.6969
177	31329	13.3041	217	47089	14.7309
178	31684	13.3417	218	47524	14.7648
179	32041	13.3791	219	47961	14.7986
180	32400	13.4164	220	48400	14.8324
181	32761	13.4536	221	48841	14.8661
182	33124	13.4907	222	49284	14.8997
183	33489	13.5277	223	49729	14.9332
184	33856	13.5647	224	50176	14.9666
185	34225	13.6015	225	50625	15.0000
186	34596	13.6382	226	51076	15.0333
187	34969	13.6748	227	51529	15.0665
188	35344	13.7113	228	51984	15.0997
189	35721	13.7477	229	52441	15.1327
190	36100	13.7840	230	52900	15.1658
191	36481	13.8203	231	53361	15.1987
192	36864	13.8564	232	53824	15.2315
193	37249	13.8924	233	54289	15.2643
194	37636	13.9284	234	54756	15.2971
195	38025	13.9642	235	55225	15.3297
196	38416	14.0000	236	55696	15.3623
197	38809	14.0357	237	56169	15.3948
198	39204	14.0712	238	56644	15.4272
199	39601	14.1067	239	57121	15.4596
200	40000	14.1421	240	57600	15.4919

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
241	58081	15.5242	281	78961	16.7631
242	58564	15.5563	282	79524	16.7929
243	59049	15.5885	283	80089	16.8226
244	59536	15.6205	284	80656	16.8523
245	60025	15.6525	285	81225	16.8819
246	60516	15.6844	286	81796	16.9115
247	61009	15.7162	287	82369	16.9411
248	61504	15.7480	288	82944	16.9706
249	62001	15.7797	289	83521	17.0000
250	62500	15.8114	290	84100	17.0294
251	63001	15.8430	291	84681	17.0587
252	63504	15.8745	292	85264	17.0880
253	64009	15.9060	293	85849	17.1172
254	64516	15.9374	294	86436	17.1464
255	65025	15.9687	295	87025	17.1756
256	65536	16.0000	296	87616	17.2047
257	66049	16.0312	297	88209	17.2337
258	66564	16.0624	298	88804	17.2627
259	67081	16.0935	299	89401	17.2916
260	67600	16.1245	300	90000	17.3205
261	68121	16.1555	301	90601	17.3494
262	68644	16.1864	302	91204	17.3781
263	69169	16.2173	303	91809	17.4069
264	69696	16.2481	304	92416	17.4356
265	70225	16.2788	305	93025	17.4642
266	70756	16.3095	306	93636	17.4929
267	71289	16.3401	307	94249	17.5214
268	71824	16.3707	308	94864	17.5499
269	72361	16.4012	309	95481	17.5784
270	72900	16.4317	310	96100	17.6068
271	73441	16.4621	311	96721	17.6352
272	73984	16.4924	312	97344	17.6635
273	74529	16.5227	313	97969	17.6918
274	75076	16.5529	314	98596	17.7200
275	75625	16.5831	315	99225	17.7482
276	76176	16.6132	316	99856	17.7764
277	76729	16.6433	317	100489	17.8045
278	77284	16.6733	318	101124	17.8326
279	77841	16.7033	319	101761	17.8606
280	78400	16.7332	320	102400	17.8885

Table 6.—Squares and Square Roots.—*(Continued)*

Number	Square	Square root	Number	Square	Square root
321	103041	17.9165	361	130321	19.0000
322	103684	17.9444	362	131044	19.0263
323	104329	17.9722	363	131769	19.0526
324	104976	18.0000	364	132496	19.0788
325	105625	18.0278	365	133225	19.1050
326	106276	18.0555	366	133956	19.1311
327	106929	18.0831	367	134689	19.1572
328	107584	18.1108	368	135424	19.1833
329	108241	18.1384	369	136161	19.2094
330	108900	18.1659	370	136900	19.2354
331	109561	18.1934	371	137641	19.2614
332	110224	18.2209	372	138384	19.2873
333	110889	18.2483	373	139129	19.3132
334	111556	18.2757	374	139876	19.3391
335	112225	18.3030	375	140625	19.3649
336	112896	18.3303	376	141376	19.3907
337	113569	18.3576	377	142129	19.4165
338	114244	18.3848	378	142884	19.4422
339	114921	18.4120	379	143641	19.4679
340	115600	18.4391	380	144400	19.4936
341	116281	18.4662	381	145161	19.5192
342	116964	18.4932	382	145924	19.5448
343	117649	18.5203	383	146689	19.5704
344	118336	18.5472	384	147456	19.5959
345	119025	18.5742	385	148225	19.6214
346	119716	18.6011	386	148996	19.6469
347	120409	18.6279	387	149769	19.6723
348	121104	18.6548	388	150544	19.6977
349	121801	18.6815	389	151321	19.7231
350	122500	18.7083	390	152100	19.7484
351	123201	18.7350	391	152881	19.7737
352	123904	18.7617	392	153664	19.7990
353	124609	18.7883	393	154449	19.8242
354	125316	18.8149	394	155236	19.8494
355	126025	18.8414	395	156025	19.8746
356	126736	18.8680	396	156816	19.8997
357	127449	18.8944	397	157609	19.9249
358	128164	18.9209	398	158404	19.9499
359	128881	18.9473	399	159201	19.9750
360	129600	18.9737	400	160000	20.0000

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
401	160801	20.0250	441	194481	21.0000
402	161604	20.0499	442	195364	21.0238
403	162409	20.0749	443	196249	21.0476
404	163216	20.0998	444	197136	21.0713
405	164025	20.1246	445	198025	21.0950
406	164836	20.1494	446	198916	21.1187
407	165649	20.1742	447	199809	21.1424
408	166464	20.1990	448	200704	21.1660
409	167281	20.2237	449	201601	21.1896
410	168100	20.2485	450	202500	21.2132
411	168921	20.2731	451	203401	21.2368
412	169744	20.2978	452	204304	21.2603
413	170569	20.3224	453	205209	21.2838
414	171396	20.3470	454	206116	21.3073
415	172225	20.3715	455	207025	21.3307
416	173056	20.3961	456	207936	21.3542
417	173889	20.4206	457	208849	21.3776
418	174724	20.4450	458	209764	21.4009
419	175561	20.4695	459	210681	21.4243
420	176400	20.4939	460	211600	21.4476
421	177241	20.5183	461	212521	21.4709
422	178084	20.5426	462	213444	21.4942
423	178929	20.5670	463	214369	21.5174
424	179776	20.5913	464	215296	21.5407
425	180625	20.6155	465	216225	21.5639
426	181476	20.6398	466	217156	21.5870
427	182329	20.6640	467	218089	21.6102
428	183184	20.6882	468	219024	21.6333
429	184041	20.7123	469	219961	21.6564
430	184900	20.7364	470	220900	21.6795
431	185761	20.7605	471	221841	21.7025
432	186624	20.7846	472	222784	21.7256
433	187489	20.8087	473	223729	21.7486
434	188356	20.8327	474	224676	21.7715
435	189225	20.8567	475	225625	21.7945
436	190096	20.8806	476	226576	21.8174
437	190969	20.9045	477	227529	21.8403
438	191844	20.9284	478	228484	21.8632
439	192721	20.9523	479	229441	21.8861
440	193600	20.9762	480	230400	21.9089

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
481	231361	21.9317	521	271441	22.8254
482	232324	21.9545	522	272484	22.8473
483	233289	21.9773	523	273529	22.8692
484	234256	22.0000	524	274576	22.8910
485	235225	22.0227	525	275625	22.9129
486	236196	22.0454	526	276676	22.9347
487	237169	22.0681	527	277729	22.9565
488	238144	22.0907	528	278784	22.9783
489	239121	22.1133	529	279841	23.0000
490	240100	22.1359	530	280900	23.0217
491	241081	22.1585	531	281961	23.0434
492	242064	22.1811	532	283024	23.0651
493	243049	22.2036	533	284089	23.0868
494	244036	22.2261	534	285156	23.1084
495	245025	22.2486	535	286225	23.1301
496	246016	22.2711	536	287296	23.1517
497	247009	22.2935	537	288369	23.1733
498	248004	22.3159	538	289444	23.1948
499	249001	22.3383	539	290521	23.2164
500	250000	22.3607	540	291600	23.2379
501	251001	22.3830	541	292681	23.2594
502	252004	22.4054	542	293764	23.2809
503	253009	22.4277	543	294849	23.3024
504	254016	22.4499	544	295936	23.3238
505	255025	22.4722	545	297025	23.3452
506	256036	22.4944	546	298116	23.3666
507	257049	22.5167	547	299209	23.3880
508	258064	22.5389	548	300304	23.4094
509	259081	22.5610	549	301401	23.4307
510	260100	22.5832	550	302500	23.4521
511	261121	22.6053	551	303601	23.4734
512	262144	22.6274	552	304704	23.4947
513	263169	22.6495	553	305809	23.5160
514	264196	22.6716	554	306916	23.5372
515	265225	22.6936	555	308025	23.5584
516	266256	22.7156	556	309136	23.5797
517	267289	22.7376	557	310249	23.6008
518	268324	22.7596	558	311364	23.6220
519	269361	22.7816	559	312481	23.6432
520	270400	22.8035	560	313600	23.6643

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
561	314721	23.6854	601	361201	24.5153
562	315844	23.7065	602	362404	24.5357
563	316969	23.7276	603	363609	24.5561
564	318096	23.7487	604	364816	24.5764
565	319225	23.7697	605	366025	24.5967
566	320356	23.7908	606	367236	24.6171
567	321489	23.8118	607	368449	24.6374
568	322624	23.8328	608	369664	24.6577
569	323761	23.8537	609	370881	24.6779
570	324900	23.8747	610	372100	24.6982
571	326041	23.8956	611	373321	24.7184
572	327184	23.9165	612	374544	24.7385
573	328329	23.9374	613	375769	24.7588
574	329476	23.9583	614	376996	24.7790
575	330625	23.9792	615	378225	24.7992
576	331776	24.0000	616	379456	24.8193
577	332929	24.0208	617	380689	24.8395
578	334084	24.0416	618	381924	24.8596
579	335241	24.0624	619	383161	24.8797
580	336400	24.0832	620	384400	24.8998
581	337561	24.1039	621	385641	24.9199
582	338724	24.1247	622	386884	24.9399
583	339889	24.1454	623	388129	24.9600
584	341056	24.1661	624	389376	24.9800
585	342225	24.1868	625	390625	25.0000
586	343396	24.2074	626	391876	25.0200
587	344569	24.2281	627	393129	25.0400
588	345744	24.2487	628	394384	25.0599
589	346921	24.2693	629	395641	25.0799
590	348100	24.2899	630	396900	25.0998
591	349281	24.3105	631	398161	25.1197
592	350464	24.3311	632	399424	25.1396
593	351649	24.3516	633	400689	25.1595
594	352836	24.3721	634	401956	25.1794
595	354025	24.3926	635	403225	25.1992
596	355216	24.4131	636	404496	25.2190
597	356409	24.4336	637	405769	25.2389
598	357604	24.4540	638	407044	25.2587
599	358801	24.4745	639	408321	25.2784
600	360000	24.4949	640	409600	25.2982

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
641	410881	25.3180	681	463761	26.0960
642	412164	25.3377	682	465124	26.1151
643	413449	25.3574	683	466489	26.1343
644	414736	25.3772	684	467856	26.1534
645	416025	25.3969	685	469225	26.1725
646	417316	25.4165	686	470596	26.1916
647	418609	25.4362	687	471969	26.2107
648	419904	25.4558	688	473344	26.2298
649	421201	25.4755	689	474721	26.2488
650	422500	25.4951	690	476100	26.2679
651	423801	25.5147	691	477481	26.2869
652	425104	25.5343	692	478864	26.3059
653	426409	25.5539	693	480249	26.3249
654	427716	25.5734	694	481636	26.3439
655	429025	25.5930	695	483025	26.3629
656	430336	25.6125	696	484416	26.3818
657	431649	25.6320	697	485809	26.4008
658	432964	25.6515	698	487204	26.4197
659	434281	25.6710	699	488601	26.4386
660	435600	25.6905	700	490000	26.4575
661	436921	25.7099	701	491401	26.4764
662	438244	25.7294	702	492804	26.4953
663	439569	25.7488	703	494209	26.5141
664	440896	25.7682	704	495616	26.5330
665	442225	25.7876	705	497025	26.5518
666	443556	25.8070	706	498436	26.5707
667	444889	25.8263	707	499849	26.5895
668	446224	25.8457	708	501264	26.6083
669	447561	25.8650	709	502681	26.6271
670	448900	25.8844	710	504100	26.6458
671	450241	25.9037	711	505521	26.6646
672	451584	25.9230	712	506944	26.6833
673	452929	25.9422	713	508369	26.7021
674	454276	25.9615	714	509796	26.7208
675	455625	25.9808	715	511225	26.7395
676	456976	26.0000	716	512656	26.7582
677	458329	26.0192	717	514089	26.7769
678	459684	26.0384	718	515524	26.7955
679	461041	26.0576	719	516961	26.8142
680	462400	26.0768	720	518400	26.8328

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
721	519841	26.8514	761	579121	27.5862
722	521284	26.8701	762	580644	27.6043
723	522729	26.8887	763	582169	27.6225
724	524176	26.9072	764	583696	27.6405
725	525625	26.9258	765	585225	27.6586
726	527076	26.9444	766	586756	27.6767
727	528529	26.9629	767	588289	27.6948
728	529984	26.9815	768	589824	27.7128
729	531441	27.0000	769	591361	27.7308
730	532900	27.0185	770	592900	27.7489
731	534361	27.0370	771	594441	27.7669
732	535824	27.0555	772	595984	27.7849
733	537289	27.0740	773	597529	27.8029
734	538756	27.0924	774	599076	27.8209
735	540225	27.1109	775	600625	27.8388
736	541696	27.1293	776	602176	27.8568
737	543169	27.1477	777	603729	27.8747
738	544644	27.1662	778	605284	27.8927
739	546121	27.1846	779	606841	27.9106
740	547600	27.2029	780	608400	27.9285
741	549081	27.2213	781	609961	27.9464
742	550564	27.2397	782	611524	27.9643
743	552049	27.2580	783	613089	27.9821
744	553536	27.2764	784	614656	28.0000
745	555025	27.2947	785	616225	28.0179
746	556516	27.3130	786	617796	28.0357
747	558009	27.3313	787	619369	28.0535
748	559504	27.3496	788	620944	28.0713
749	561001	27.3679	789	622521	28.0891
750	562500	27.3861	790	624100	28.1069
751	564001	27.4044	791	625681	28.1247
752	565504	27.4226	792	627264	28.1425
753	567009	27.4408	793	628849	28.1603
754	568516	27.4591	794	630436	28.1780
755	570025	27.4773	795	632025	28.1957
756	571536	27.4955	796	633616	28.2135
757	573049	27.5136	797	635209	28.2312
758	574564	27.5318	798	636804	28.2489
759	576081	27.5500	799	638401	28.2666
760	577600	27.5681	800	640000	28.2843

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
801	641601	28.3019	841	707281	29.0000
802	643204	28.3196	842	708964	29.0172
803	644809	28.3373	843	710649	29.0345
804	646416	28.3549	844	712336	29.0517
805	648025	28.3725	845	714025	29.0689
806	649636	28.3901	846	715716	29.0861
807	651249	28.4077	847	717409	29.1033
808	652864	28.4253	848	719104	29.1204
809	654481	28.4429	849	720801	29.1376
810	656100	28.4605	850	722500	29.1548
811	657721	28.4781	851	724201	29.1719
812	659344	28.4956	852	725904	29.1890
813	660969	28.5132	853	727609	29.2062
814	662596	28.5307	854	729316	29.2233
815	664225	28.5482	855	731025	29.2404
816	665856	28.5657	856	732736	29.2575
817	667489	28.5832	857	734449	29.2746
818	669124	28.6007	858	736164	29.2916
819	670761	28.6182	859	737881	29.3087
820	672400	28.6356	860	739600	29.3258
821	674041	28.6531	861	741321	29.3428
822	675684	28.6705	862	743044	29.3598
823	677329	28.6880	863	744769	29.3769
824	678976	28.7054	864	746496	29.3939
825	680625	28.7228	865	748225	29.4109
826	682276	28.7402	866	749956	29.4279
827	683929	28.7576	867	751689	29.4449
828	685584	28.7750	868	753424	29.4618
829	687241	28.7924	869	755161	29.4788
830	688900	28.8097	870	756900	29.4958
831	690561	28.8271	871	758641	29.5127
832	692224	28.8444	872	760384	29.5296
833	693889	28.8617	873	762129	29.5466
834	695556	28.8791	874	763876	29.5635
835	697225	28.8964	875	765625	29.5804
836	698896	28.9137	876	767376	29.5973
837	700569	28.9310	877	769129	29.6142
838	702244	28.9482	878	770884	29.6311
839	703921	28.9655	879	772641	29.6479
840	705600	28.9828	880	774400	29.6648

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
881	776161	29.6816	921	848241	30.3480
882	777924	29.6985	922	850084	30.3645
883	779689	29.7153	923	851929	30.3809
884	781456	29.7321	924	853776	30.3974
885	783225	29.7489	925	855625	30.4138
886	784996	29.7658	926	857476	30.4302
887	786769	29.7825	927	859329	30.4467
888	788544	29.7993	928	861184	30.4631
889	790321	29.8161	929	863041	30.4795
890	792100	29.8329	930	864900	30.4959
891	793881	29.8496	931	866761	30.5123
892	795664	29.8664	932	868624	30.5287
893	797449	29.8831	933	870489	30.5450
894	799236	29.8998	934	872356	30.5614
895	801025	29.9166	935	874225	30.5778
896	802816	29.9333	936	876096	30.5941
897	804609	29.9500	937	877969	30.6105
898	806404	29.9666	938	879844	30.6268
899	808201	29.9833	939	881721	30.6431
900	810000	30.0000	940	883600	30.6594
901	811801	30.0167	941	885481	30.6757
902	813604	30.0333	942	887364	30.6920
903	815409	30.0500	943	889249	30.7083
904	817216	30.0666	944	891136	30.7246
905	819025	30.0832	945	893025	30.7409
906	820836	30.0998	946	894916	30.7571
907	822649	30.1164	947	896809	30.7734
908	824464	30.1330	948	898704	30.7896
909	826281	30.1496	949	900601	30.8058
910	828100	30.1662	950	902500	30.8221
911	829921	30.1828	951	904401	30.8383
912	831744	30.1993	952	906304	30.8545
913	833569	30.2159	953	908209	30.8707
914	835396	30.2324	954	910116	30.8869
915	837225	30.2490	955	912025	30.9031
916	839056	30.2655	956	913936	30.9192
917	840889	30.2820	957	915849	30.9354
918	842724	30.2985	958	917764	30.9516
919	844561	30.3150	959	919681	30.9677
920	846400	30.3315	960	921600	30.9839

TABLE 6.—SQUARES AND SQUARE ROOTS.—(*Continued*)

Number	Square	Square root	Number	Square	Square root
961	923521	31.0000	981	962361	31.3209
962	925444	31.0161	982	964324	31.3369
963	927369	31.0322	983	966289	31.3528
964	929296	31.0483	984	968256	31.3688
965	931225	31.0644	985	970225	31.3847
966	933156	31.0805	986	972196	31.4006
967	935089	31.0966	987	974169	31.4166
968	937024	31.1127	988	976144	31.4325
969	938961	31.1288	989	978121	31.4484
970	940900	31.1448	990	980100	31.4643
971	942841	31.1609	991	982081	31.4802
972	944784	31.1769	992	984064	31.4960
973	946729	31.1929	993	986049	31.5119
974	948676	31.2090	994	988036	31.5278
975	950625	31.2250	995	990025	31.5436
976	952576	31.2410	996	992016	31.5595
977	954529	31.2570	997	994009	31.5753
978	956484	31.2730	998	996004	31.5911
979	958441	31.2890	999	998001	31.6070
980	960400	31.3050	1000	1000000	31.6228

Foreword to Table 7.—Logarithms are the greatest labor-saving discovery ever made in the field of mathematics. With their aid, many calculations can be performed easily and quickly that would not be feasible at all without them.

The common logarithm of a number is the power to which 10 must be raised to produce that number. For example, $10^2 = 100$, so the logarithm of 100 is 2. Similarly, $90 = 10^{1.95424}$, and the logarithm of 90 is 1.95424. In general, if $Y = 10^x$, then $\log Y = x$.

That part of the logarithm to the left of the decimal is called the characteristic, while that part to the right of the decimal is the mantissa. Thus, for $\log 10 = 1.95424$, the characteristic is 1, the mantissa is .95424.

There are three fundamental principles that are constantly needed in working with logarithms:

1. $\log (ab) = \log a + \log b$,
2. $\log (a/b) = \log a - \log b$,
3. $\log (a^n) = n \log a$.

To find the mantissa of any number, we enter Table 7, find the first three digits of the number in the left-hand column, headed "No.," and find the fourth digit in the top row, then read off the mantissa from the proper row and column. Thus, for the number 1,503, we find 150 in the first column and "3" in the fifth column of the table, and read off the mantissa .17696.

The characteristic of a logarithm is discovered by placing the pencil point between the first two significant figures (the first figure that is not zero is the first significant figure) of the number, and moving it to the right or left so many places to the decimal point. If the pencil is moved to the right, the characteristic is positive; if to the left, it is negative. Thus, for the number 1,503, the pencil is placed between 1 and 5, and moved three places to the right. The characteristic is therefore 3, and the complete logarithm is 3.17696.

If the number is 15,030, the mantissa is the same, but the characteristic is 4, so that the logarithm is 4.17696.

In the case of the number 15,037, the exact mantissa cannot be read directly from Table 7, but an approximate mantissa can be obtained by taking the mantissa of the number 1,504, or, more accurately, by interpolating. To interpolate, we subtract the mantissa of the number just smaller than the given number from the mantissa of the number just larger, then subtract from the given number the table number just smaller, place a decimal point before the first figure of this last difference, multiply the first difference by this value, and add the product to the mantissa of the table number just smaller than the given number. Thus, the mantissa of 15,030 is .17696, the mantissa of 15,040 is .17725, and their difference is .00029; the difference between the given number and the table number just smaller is $15,037 - 15,030 = 7$, which becomes .7; the product of the first difference and this value is $.00029 \times .7 = .000203$; and this product added to the mantissa of 15,030 is $.17696 + .000203 = .17716$. The logarithm of 15,037 is therefore 4.17716.

Suppose that the number whose logarithm is required is 15.037. The mantissa is the same as that just found for the number 15,037, but the

characteristic changes from 4 to 1, so that the logarithm is now 1.17716. Similarly, the logarithm of 1.5037 is 0.17716.

If we move the decimal one or more places to the left, giving, say, the number .15037, we do not change the mantissa, but we encounter a negative characteristic. For, if we put a pencil between 1 and 5, we must move it one place to the left to reach the decimal point. The characteristic is then −1. To avoid these awkward negative characteristics, it is customary to write −1 in the form 9 . . . . − 10, −2 in the form 8 . . . . −10, etc. Hence, the logarithm of .15037 is written 9.17716 − 10.

To obtain from Table 7 the number corresponding to a logarithm, we find in the table the mantissa of the logarithm, write down the number corresponding to it, and then point off this number in accordance with the characteristic of the logarithm. For example, given the logarithm 2.27921, we look in the table for the mantissa .27921, read off the corresponding number 1,902, and point off this number by placing our pencil point between the figures 1 and 9, then moving it two places to the right as indicated by the positive characteristic, 2, getting 190.2 as the result. If the logarithm is 0.27921, the number is 1.902; if the logarithm is 8.27921 − 10, the number is 0.01902; and so on.

Let us now find a geometric mean by the use of logarithms. By formula (14) of Chap. VII,

$$G = (5 \cdot 11 \cdot 19)^{\frac{1}{3}*}.$$

According to principle 3, above,

$$\log G = \tfrac{1}{3} \log (5 \cdot 11 \cdot 19).$$

And by principle 1,

$$\log G = \tfrac{1}{3} (\log 5 + \log 11 + \log 19).$$

Now, the numbers 5, 11, and 19 do not appear in Table 7, but the numbers 5,000, 1,100, and 1,900, which have the same mantissas, may be found there. The mantissa of 5,000 is .69897, and the characteristic of 5 is 0, so log 5 is 0.69897. In the same way, we find log 11 = 1.04139, and log 19 = 1.27875. We therefore have

$$\log G = \tfrac{1}{3} (0.69897 + 1.04139 + 1.27875) = \tfrac{1}{3} (3.01911)$$
$$= 1.00637.$$

Looking in the table for the mantissa .00637, the nearest we can find to it is the mantissa .00647, to which the corresponding number is 1,015. Pointing this off according to the characteristic, 1, we get 10.15 as the geometric mean. If greater accuracy is wanted, we may interpolate in the table. Our mantissa, .00637, falls between the two tabular mantissas .00604 and .00647. We therefore have .00637 − .00604 = .00033; .00647 − .00604 = .00043; and .00033/.00043 = .767. That is, our matissa indicates a number about $\frac{3}{4}$ of the way between 10.14 and 10.15, or roughly 10.14 + .00767 = 10.14767.

* $5 \cdot 11 \cdot 19 = 5 \times 11 \times 19$.

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS
100–149

No.	0	1	2	3	4	5	6	7	8	9
100	00 000	00 043	00 087	00 130	00 173	00 217	00 260	00 303	00 346	00 389
101	00 432	00 475	00 518	00 561	00 604	00 647	00 689	00 732	00 775	00 817
102	00 860	00 903	00 945	00 988	01 030	01 072	01 115	01 157	01 199	01 242
103	01 284	01 326	01 368	01 410	01 452	01 494	01 536	01 578	01 620	01 662
104	01 703	01 745	01 787	01 828	01 870	01 912	01 953	01 995	02 036	02 078
105	02 119	02 160	02 202	02 243	02 284	02 325	02 366	02 407	02 449	02 490
106	02 531	02 572	02 612	02 653	02 694	02 735	02 776	02 816	02 857	02 898
107	02 938	02 979	03 019	03 060	03 100	03 141	03 181	03 222	03 262	03 302
108	03 342	03 383	03 423	03 463	03 503	03 543	03 583	03 623	03 663	03 703
109	03 743	03 782	03 822	03 862	03 902	03 941	03 981	04 021	04 060	04 100
110	04 139	04 179	04 218	04 258	04 297	04 336	04 376	04 415	04 454	04 493
111	04 532	04 571	04 610	04 650	04 689	04 727	04 766	04 805	04 844	04 883
112	04 922	04 961	04 999	05 038	05 077	05 115	05 154	05 192	05 231	05 269
113	05 308	05 346	05 385	05 423	05 461	05 500	05 538	05 576	05 614	05 652
114	05 690	05 729	05 767	05 805	05 843	05 881	05 918	05 956	05 994	06 032
115	06 070	06 108	06 145	06 183	06 221	06 258	06 296	06 333	06 371	06 408
116	06 446	06 483	06 521	06 558	06 595	06 633	06 670	06 707	06 744	06 781
117	06 819	06 856	06 893	06 930	06 967	07 004	07 041	07 078	07 115	07 151
118	07 188	07 225	07 262	07 298	07 335	07 372	07 408	07 445	07 482	07 518
119	07 555	07 591	07 628	07 664	07 700	07 737	07 773	07 809	07 846	07 882
120	07 918	07 954	07 990	08 027	08 063	08 099	08 135	08 171	08 207	08 243
121	08 279	08 314	08 350	08 386	08 422	08 458	08 493	08 529	08 565	08 600
122	08 636	08 672	08 707	08 743	08 778	08 814	08 849	08 884	08 920	08 955
123	08 991	09 026	09 061	09 096	09 132	09 167	09 202	09 237	09 272	09 307
124	09 342	09 377	09 412	09 447	09 482	09 517	09 552	09 587	09 621	09 656
125	09 691	09 726	09 760	09 795	09 830	09 864	09 899	09 934	09 968	10 003
126	10 037	10 072	10 106	10 140	10 175	10 209	10 243	10 278	10 312	10 346
127	10 380	10 415	10 449	10 483	10 517	10 551	10 585	10 619	10 653	10 687
128	10 721	10 755	10 789	10 823	10 857	10 890	10 924	10 958	10 992	11 025
129	11 059	11 093	11 126	11 160	11 193	11 227	11 261	11 294	11 327	11 361
130	11 394	11 428	11 461	11 494	11 528	11 561	11 594	11 628	11 661	11 694
131	11 727	11 760	11 793	11 826	11 860	11 893	11 926	11 959	11 992	12 024
132	12 057	12 090	12 123	12 156	12 189	12 222	12 254	12 287	12 320	12 352
133	12 385	12 418	12 450	12 483	12 516	12 548	12 581	12 613	12 646	12 678
134	12 710	12 743	12 775	12 808	12 840	12 872	12 905	12 937	12 969	13 001
135	13 033	13 066	13 098	13 130	13 162	13 194	13 226	13 258	13 290	13 322
136	13 354	13 386	13 418	13 450	13 481	13 513	13 545	13 577	13 609	13 640
137	13 672	13 704	13 735	13 767	13 799	13 830	13 862	13 893	13 925	13 956
138	13 988	14 019	14 051	14 082	14 114	14 145	14 176	14 208	14 239	14 270
139	14 301	14 333	14 364	14 395	14 426	14 457	14 489	14 520	14 551	14 582
140	14 613	14 644	14 675	14 706	14 737	14 768	14 799	14 829	14 860	14 891
141	14 922	14 953	14 983	15 014	15 045	15 076	15 106	15 137	15 168	15 198
142	15 229	15 259	15 290	15 320	15 351	15 381	15 412	15 442	15 473	15 503
143	15 534	15 564	15 594	15 625	15 655	15 685	15 715	15 746	15 776	15 806
144	15 836	15 866	15 897	15 927	15 957	15 987	16 017	16 047	16 077	16 107
145	16 137	16 167	16 197	16 227	16 256	16 286	16 316	16 346	16 376	16 406
146	16 435	16 465	16 495	16 524	16 554	16 584	16 613	16 643	16 673	16 702
147	16 732	16 761	16 791	16 820	16 850	16 879	16 909	16 938	16 967	16 997
148	17 026	17 056	17 085	17 114	17 143	17 173	17 202	17 231	17 260	17 289
149	17 319	17 348	17 377	17 406	17 435	17 464	17 493	17 522	17 551	17 580
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)
150–199

No.	0	1	2	3	4	5	6	7	8	9
150	17 609	17 638	17 667	17 696	17 725	17 754	17 782	17 811	17 840	17 869
151	17 898	17 926	17 955	17 984	18 013	18 041	18 070	18 099	18 127	18 156
152	18 184	18 213	18 241	18 270	18 298	18 327	18 355	18 384	18 412	18 441
153	18 469	18 498	18 526	18 554	18 583	18 611	18 639	18 667	18 696	18 724
154	18 752	18 780	18 808	18 837	18 865	18 893	18 921	18 949	18 977	19 005
155	19 033	19 061	19 089	19 117	19 145	19 173	19 201	19 229	19 257	19 285
156	19 312	19 340	19 368	19 396	19 424	19 451	19 479	19 507	19 535	19 562
157	19 590	19 618	19 645	19 673	19 700	19 728	19 756	19 783	19 811	19 838
158	19 866	19 893	19 921	19 948	19 976	20 003	20 030	20 058	20 085	20 112
159	20 140	20 167	20 194	20 222	20 249	20 276	20 303	20 330	20 358	20 385
160	20 412	20 439	20 466	20 493	20 520	20 548	20 575	20 602	20 629	20 656
161	20 683	20 710	20 737	20 763	20 790	20 817	20 844	20 871	20 898	20 925
162	20 952	20 978	21 005	21 032	21 059	21 085	21 112	21 139	21 165	21 192
163	21 219	21 245	21 272	21 299	21 325	21 352	21 378	21 405	21 431	21 458
164	21 484	21 511	21 537	21 564	21 590	21 617	21 643	21 669	21 696	21 722
165	21 748	21 775	21 801	21 827	21 854	21 880	21 906	21 932	21 958	21 985
166	22 011	22 037	22 063	22 089	22 115	22 141	22 167	22 194	22 220	22 246
167	22 272	22 298	22 324	22 350	22 376	22 401	22 427	22 453	22 479	22 505
168	22 531	22 557	22 583	22 608	22 634	22 660	22 686	22 712	22 737	22 763
169	22 789	22 814	22 840	22 866	22 891	22 917	22 943	22 968	22 994	23 019
170	23 045	23 070	23 096	23 121	23 147	23 172	23 198	23 223	23 249	23 274
171	23 300	23 325	23 350	23 376	23 401	23 426	23 452	23 477	23 502	23 528
172	23 553	23 578	23 603	23 629	23 654	23 679	23 704	23 729	23 754	23 779
173	23 805	23 830	23 855	23 880	23 905	23 930	23 955	23 980	24 005	24 030
174	24 055	24 080	24 105	24 130	24 155	24 180	24 204	24 229	24 254	24 279
175	24 304	24 329	24 353	24 378	24 403	24 428	24 452	24 477	24 502	24 527
176	24 551	24 576	24 601	24 625	24 650	24 674	24 699	24 724	24 748	24 773
177	24 797	24 822	24 846	24 871	24 895	24 920	24 944	24 969	24 993	25 018
178	25 042	25 066	25 091	25 115	25 139	25 164	25 188	25 212	25 237	25 261
179	25 285	25 310	25 334	25 358	25 382	25 406	25 431	25 455	25 479	25 503
180	25 527	25 551	25 575	25 600	25 624	25 648	25 672	25 696	25 720	25 744
181	25 768	25 792	25 816	25 840	25 864	25 888	25 912	25 935	25 959	25 983
182	26 007	26 031	26 055	26 079	26 102	26 126	26 150	26 174	26 198	26 221
183	26 245	26 269	26 293	26 316	26 340	26 364	26 387	26 411	26 435	26 458
184	26 482	26 505	26 529	26 553	26 576	26 600	26 623	26 647	26 670	26 694
185	26 717	26 741	26 764	26 788	26 811	26 834	26 858	26 881	26 905	26 928
186	26 951	26 975	26 998	27 021	27 045	27 068	27 091	27 114	27 138	27 161
187	27 184	27 207	27 231	27 254	27 277	27 300	27 323	27 346	27 370	27 393
188	27 416	27 439	27 462	27 485	27 508	27 531	27 554	27 577	27 600	27 623
189	27 646	27 669	27 692	27 715	27 738	27 761	27 784	27 807	27 830	27 852
190	27 875	27 898	27 921	27 944	27 967	27 989	28 012	28 035	28 058	28 081
191	28 103	28 126	28 149	28 171	28 194	28 217	28 240	28 262	28 285	28 307
192	28 330	28 353	28 375	28 398	28 421	28 443	28 466	28 488	28 511	28 533
193	28 556	28 578	28 601	28 623	28 646	28 668	28 691	28 713	28 735	28 758
194	28 780	28 803	28 825	28 847	28 870	28 892	28 914	28 937	28 959	28 981
195	29 003	29 026	29 048	29 070	29 092	29 115	29 137	29 159	29 181	29 203
196	29 226	29 248	29 270	29 292	29 314	29 336	29 358	29 380	29 403	29 425
197	29 447	29 469	29 491	29 513	29 535	29 557	29 579	29 601	29 623	29 645
198	29 667	29 688	29 710	29 732	29 754	29 776	29 798	29 820	29 842	29 863
199	29 885	29 907	29 929	29 951	29 973	29 994	30 016	30 038	30 060	30 081
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

200–249

No.	0	1	2	3	4	5	6	7	8	9
200	30 103	30 125	30 146	30 168	30 190	30 211	30 233	30 255	30 276	30 298
201	30 320	30 341	30 363	30 384	30 406	30 428	30 449	30 471	30 492	30 514
202	30 535	30 557	30 578	30 600	30 621	30 643	30 664	30 685	30 707	30 728
203	30 750	30 771	30 792	30 814	30 835	30 856	30 878	30 899	30 920	30 942
204	30 963	30 984	31 006	31 027	31 048	31 069	31 091	31 112	31 133	31 154
205	31 175	31 197	31 218	31 239	31 260	31 281	31 302	31 323	31 345	31 366
206	31 387	31 408	31 429	31 450	31 471	31 492	31 513	31 534	31 555	31 576
207	31 597	31 618	31 639	31 660	31 681	31 702	31 723	31 744	31 765	31 785
208	31 806	31 827	31 848	31 869	31 890	31 911	31 931	31 952	31 973	31 994
209	32 015	32 035	32 056	32 077	32 098	32 118	32 139	32 160	32 181	32 201
210	32 222	32 243	32 263	32 284	32 305	32 325	32 346	32 366	32 387	32 408
211	32 428	32 449	32 469	32 490	32 510	32 531	32 552	32 572	32 593	32 613
212	32 634	32 654	32 675	32 695	32 715	32 736	32 756	32 777	32 797	32 818
213	32 838	32 858	32 879	32 899	32 919	32 940	32 960	32 980	33 001	33 021
214	33 041	33 062	33 082	33 102	33 122	33 143	33 163	33 183	33 203	33 224
215	33 244	33 264	33 284	33 304	33 325	33 345	33 365	33 385	33 405	33 425
216	33 445	33 465	33 486	33 506	33 526	33 546	33 566	33 586	33 606	33 626
217	33 646	33 666	33 686	33 706	33 726	33 746	33 766	33 786	33 806	33 826
218	33 846	33 866	33 885	33 905	33 925	33 945	33 965	33 985	34 005	34 025
219	34 044	34 064	34 084	34 104	34 124	34 143	34 163	34 183	34 203	34 223
220	34 242	34 262	34 282	34 301	34 321	34 341	34 361	34 380	34 400	34 420
221	34 439	34 459	34 479	34 498	34 518	34 537	34 557	34 577	34 596	34 616
222	34 635	34 655	34 674	34 694	34 713	34 733	34 753	34 772	34 792	34 811
223	34 830	34 850	34 869	34 889	34 908	34 928	34 947	34 967	34 986	35 005
224	35 025	35 044	35 064	35 083	35 102	35 122	35 141	35 160	35 180	35 199
225	35 218	35 238	35 257	35 276	35 295	35 315	35 334	35 353	35 372	35 392
226	35 411	35 430	35 449	35 468	35 488	35 507	35 526	35 545	35 564	35 583
227	35 603	35 622	35 641	35 660	35 679	35 698	35 717	35 736	35 755	35 774
228	35 793	35 813	35 832	35 851	35 870	35 889	35 908	35 927	35 946	35 965
229	35 984	36 003	36 021	36 040	36 059	36 078	36 097	36 116	36 135	36 154
230	36 173	36 192	36 211	36 229	36 248	36 267	36 286	36 305	36 324	36 342
231	36 361	36 380	36 399	36 418	36 436	36 455	36 474	36 493	36 511	36 530
232	36 549	36 568	36 586	36 605	36 624	36 642	36 661	36 680	36 698	36 717
233	36 736	36 754	36 773	36 791	36 810	36 829	36 847	36 866	36 884	36 903
234	36 922	36 940	36 959	36 977	36 996	37 014	37 033	37 051	37 070	37 088
235	37 107	37 125	37 144	37 162	37 181	37 199	37 218	37 236	37 254	37 273
236	37 291	37 310	37 328	37 346	37 365	37 383	37 401	37 420	37 438	37 457
237	37 475	37 493	37 511	37 530	37 548	37 566	37 585	37 603	37 621	37 639
238	37 658	37 676	37 694	37 712	37 731	37 749	37 767	37 785	37 803	37 822
239	37 840	37 858	37 876	37 894	37 912	37 931	37 949	37 967	37 985	38 003
240	38 021	38 039	38 057	38 075	38 093	38 112	38 130	38 148	38 166	38 184
241	38 202	38 220	38 238	38 256	38 274	38 292	38 310	38 328	38 346	38 364
242	38 382	38 399	38 417	38 435	38 453	38 471	38 489	38 507	38 525	38 543
243	38 561	38 578	38 596	38 614	38 632	38 650	38 668	38 686	38 703	38 721
244	38 739	38 757	38 775	38 792	38 810	38 828	38 846	38 863	38 881	38 899
245	38 917	38 934	38 952	38 970	38 987	39 005	39 023	39 041	39 058	39 076
246	39 094	39 111	39 129	39 146	39 164	39 182	39 199	39 217	39 235	39 252
247	39 270	39 287	39 305	39 322	39 340	39 358	39 375	39 393	39 410	39 428
248	39 445	39 463	39 480	39 498	39 515	39 533	39 550	39 568	39 585	39 602
249	39 620	39 637	39 655	39 672	39 690	39 707	39 724	39 742	39 759	39 777
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(Continued)
250–299

No.	0	1	2	3	4	5	6	7	8	9
250	39 794	39 811	39 829	39 846	39 863	39 881	39 898	39 915	39 933	39 950
251	39 967	39 985	40 002	40 019	40 037	40 054	40 071	40 088	40 106	40 123
252	40 140	40 157	40 175	40 192	40 209	40 226	40 243	40 261	40 278	40 295
253	40 312	40 329	40 346	40 364	40 381	40 398	40 415	40 432	40 449	40 466
254	40 483	40 500	40 518	40 535	40 552	40 569	40 586	40 603	40 620	40 637
255	40 654	40 671	40 688	40 705	40 722	40 739	40 756	40 773	40 790	40 807
256	40 824	40 841	40 858	40 875	40 892	40 909	40 926	40 943	40 960	40 976
257	40 993	41 010	41 027	41 044	41 061	41 078	41 095	41 111	41 128	41 145
258	41 162	41 179	41 196	41 212	41 229	41 246	41 263	41 280	41 296	41 313
259	41 330	41 347	41 363	41 380	41 397	41 414	41 430	41 447	41 464	41 481
260	41 497	41 514	41 531	41 547	41 564	41 581	41 597	41 614	41 631	41 647
261	41 664	41 681	41 697	41 714	41 731	41 747	41 764	41 780	41 797	41 814
262	41 830	41 847	41 863	41 880	41 896	41 913	41 929	41 946	41 963	41 979
263	41 996	42 012	42 029	42 045	42 062	42 078	42 095	42 111	42 127	42 144
264	42 160	42 177	42 193	42 210	42 226	42 243	42 259	42 275	42 292	42 308
265'	42 325	42 341	42 357	42 374	42 390	42 406	42 423	42 439	42 455	42 472
266	42 488	42 504	42 521	42 537	42 553	42 570	42 586	42 602	42 619	42 635
267	42 651	42 667	42 684	42 700	42 716	42 732	42 749	42 765	42 781	42 797
268	42 813	42 830	42 846	42 862	42 878	42 894	42 911	42 927	42 943	42 959
269	42 975	42 991	43 008	43 024	43 040	43 056	43 072	43 088	43 104	43 120
270	43 136	43 152	43 169	43 185	43 201	43 217	43 233	43 249	43 265	43 281
271	43 297	43 313	43 329	43 345	43 361	43 377	43 393	43 409	43 425	43 441
272	43 457	43 473	43 489	43 505	43 521	43 537	43 553	43 569	43 584	43 600
273	43 616	43 632	43 648	43 664	43 680	43 696	43 712	43 727	43 743	43 759
274	43 775	43 791	43 807	43 823	43 838	43 854	43 870	43 886	43 902	43 917
275	43 933	43 949	43 965	43 981	43 996	44 012	44 028	44 044	44 059	44 075
276	44 091	44 107	44 122	44 138	44 154	44 170	44 185	44 201	44 217	44 232
277	44 248	44 264	44 279	44 295	44 311	44 326	44 342	44 358	44 373	44 389
278	44 404	44 420	44 436	44 451	44 467	44 483	44 498	44 514	44 529	44 545
279	44 560	44 576	44 592	44 607	44 623	44 638	44 654	44 669	44 685	44 700
280	44 716	44 731	44 747	44 762	44 778	44 793	44 809	44 824	44 840	44 855
281	44 871	44 886	44 902	44 917	44 932	44 948	44 963	44 979	44 994	45 010
282	45 025	45 040	45 056	45 071	45 086	45 102	45 117	45 133	45 148	45 163
283	45 179	45 194	45 209	45 225	45 240	45 255	45 271	45 286	45 301	45 317
284	45 332	45 347	45 362	45 378	45 393	45 408	45 423	45 439	45 454	45 469
285	45 484	45 500	45 515	45 530	45 545	45 561	45 576	45 591	45 606	45 621
286	45 637	45 652	45 667	45 682	45 697	45 712	45 728	45 743	45 758	45 773
287	45 788	45 803	45 818	45 834	45 849	45 864	45 879	45 894	45 909	45 924
288	45 939	45 954	45 969	45 984	46 000	46 015	46 030	46 045	46 060	46 075
289	46 090	46 105	46 120	46 135	46 150	46 165	46 180	46 195	46 210	46 225
290	46 240	46 255	46 270	46 285	46 300	46 315	46 330	46 345	46 359	46 374
291	46 389	46 404	46 419	46 434	46 449	46 464	46 479	46 494	46 509	46 523
292	46 538	46 553	46 568	46 583	46 598	46 613	46 627	46 642	46 657	46 672
293	46 687	46 702	46 716	46 731	46 746	46 761	46 776	46 790	46 805	46 820
294	46 835	46 850	46 864	46 879	46 894	46 909	46 923	46 938	46 953	46 967
295	46 982	46 997	47 012	47 026	47 041	47 056	47 070	47 085	47 100	47 114
296	47 129	47 144	47 159	47 173	47 188	47 202	47 217	47 232	47 246	47 261
297	47 276	47 290	47 305	47 319	47 334	47 349	47 363	47 378	47 392	47 407
298	47 422	47 436	47 451	47 465	47 480	47 494	47 509	47 524	47 538	47 553
299	47 567	47 582	47 596	47 611	47 625	47 640	47 654	47 669	47 683	47 698
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—*(Continued)*

300–349

No.	0	1	2	3	4	5	6	7	8	9
300	47 712	47 727	47 741	47 756	47 770	47 784	47 799	47 813	47 828	47 842
301	47 857	47 871	47 885	47 900	47 914	47 929	47 943	47 958	47 972	47 986
302	48 001	48 015	48 029	48 044	48 058	48 073	48 087	48 101	48 116	48 130
303	48 144	48 159	48 173	48 187	48 202	48 216	48 230	48 244	48 259	48 273
304	48 287	48 302	48 316	48 330	48 344	48 359	48 373	48 387	48 401	48 416
305	48 430	48 444	48 458	48 473	48 487	48 501	48 515	48 530	48 544	48 558
306	48 572	48 586	48 601	48 615	48 629	48 643	48 657	48 671	48 686	48 700
307	48 714	48 728	48 742	48 756	48 770	48 785	48 799	48 813	48 827	48 841
308	48 855	48 869	48 883	48 897	48 911	48 926	48 940	48 954	48 968	48 982
309	48 996	49 010	49 024	49 038	49 052	49 066	49 080	49 094	49 108	49 122
310	49 136	49 150	49 164	49 178	49 192	49 206	49 220	49 234	49 248	49 262
311	49 276	49 290	49 304	49 318	49 332	49 346	49 360	49 374	49 388	49 402
312	49 415	49 429	49 443	49 457	49 471	49 485	49 499	49 513	49 527	49 541
313	49 554	49 568	49 582	49 596	49 610	49 624	49 638	49 651	49 665	49 679
314	49 693	49 707	49 721	49 734	49 748	49 762	49 776	49 790	49 803	49 817
315	49 831	49 845	49 859	49 872	49 886	49 900	49 914	49 927	49 941	49 955
316	49 969	49 982	49 996	50 010	50 024	50 037	50 051	50 065	50 079	50 092
317	50 106	50 120	50 133	50 147	50 161	50 174	50 188	50 202	50 215	50 229
318	50 243	50 256	50 270	50 284	50 297	50 311	50 325	50 338	50 352	50 365
319	50 379	50 393	50 406	50 420	50 433	50 447	50 461	50 474	50 488	50 501
320	50 515	50 529	50 542	50 556	50 569	50 583	50 596	50 610	50 623	50 637
321	50 651	50 664	50 678	50 691	50 705	50 718	50 732	50 745	50 759	50 772
322	50 786	50 799	50 813	50 826	50 840	50 853	50 866	50 880	50 893	50 907
323	50 920	50 934	50 947	50 961	50 974	50 987	51 001	51 014	51 028	51 041
324	51 055	51 068	51 081	51 095	51 108	51 121	51 135	51 148	51 162	51 175
325	51 188	51 202	51 215	51 228	51 242	51 255	51 268	51 282	51 295	51 308
326	51 322	51 335	51 348	51 362	51 375	51 388	51 402	51 415	51 428	51 441
327	51 455	51 468	51 481	51 495	51 508	51 521	51 534	51 548	51 561	51 574
328	51 587	51 601	51 614	51 627	51 640	51 654	51 667	51 680	51 693	51 706
329	51 720	51 733	51 746	51 759	51 772	51 786	51 799	51 812	51 825	51 838
330	51 851	51 865	51 878	51 891	51 904	51 917	51 930	51 943	51 957	51 970
331	51 983	51 996	52 009	52 022	52 035	52 048	52 061	52 075	52 088	52 101
332	52 114	52 127	52 140	52 153	52 166	52 179	52 192	52 205	52 218	52 231
333	52 244	52 257	52 270	52 284	52 297	52 310	52 323	52 336	52 349	52 362
334	52 375	52 388	52 401	52 414	52 427	52 440	52 453	52 466	52 479	52 492
335	52 504	52 517	52 530	52 543	52 556	52 569	52 582	52 595	52 608	52 621
336	52 634	52 647	52 660	52 673	52 686	52 699	52 711	52 724	52 737	52 750
337	52 763	52 776	52 789	52 802	52 815	52 827	52 840	52 853	52 866	52 879
338	52 892	52 905	52 917	52 930	52 943	52 956	52 969	52 982	52 994	53 007
339	53 020	53 033	53 046	53 058	53 071	53 084	53 097	53 110	53 122	53 135
340	53 148	53 161	53 173	53 186	53 199	53 212	53 224	53 237	53 250	53 263
341	53 275	53 288	53 301	53 314	53 326	53 339	53 352	53 364	53 377	53 390
342	53 403	53 415	53 428	53 441	53 453	53 466	53 479	53 491	53 504	53 517
343	53 529	53 542	53 555	53 567	53 580	53 593	53 605	53 618	53 631	53 643
344	53 656	53 668	53 681	53 694	53 706	53 719	53 732	53 744	53 757	53 769
345	53 782	53 794	53 807	53 820	53 832	53 845	53 857	53 870	53 882	53 895
346	53 908	53 920	53 933	53 945	53 958	53 970	53 983	53 995	54 008	54 020
347	54 033	54 045	54 058	54 070	54 083	54 095	54 108	54 120	54 133	54 145
348	54 158	54 170	54 183	54 195	54 208	54 220	54 233	54 245	54 258	54 270
349	54 283	54 295	54 307	54 320	54 332	54 345	54 357	54 370	54 382	54 394
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—*(Continued)*
350–399

No.	0	1	2	3	4	5	6	7	8	9
350	54 407	54 419	54 432	54 444	54 456	54 469	54 481	54 494	54 506	54 518
351	54 531	54 543	54 555	54 568	54 580	54 593	54 605	54 617	54 630	54 642
352	54 654	54 667	54 679	54 691	54 704	54 716	54 728	54 741	54 753	54 765
353	54 777	54 790	54 802	54 814	54 827	54 839	54 851	54 864	54 876	54 888
354	54 900	54 913	54 925	54 937	54 949	54 962	54 974	54 986	54 998	55 011
355	55 023	55 035	55 047	55 060	55 072	55 084	55 096	55 108	55 121	55 133
356	55 145	55 157	55 169	55 182	55 194	55 206	55 218	55 230	55 242	55 255
357	55 267	55 279	55 291	55 303	55 315	55 328	55 340	55 352	55 364	55 376
358	55 388	55 400	55 413	55 425	55 437	55 449	55 461	55 473	55 485	55 497
359	55 509	55 522	55 534	55 546	55 558	55 570	55 582	55 594	55 606	55 618
360	55 630	55 642	55 654	55 666	55 678	55 691	55 703	55 715	55 727	55 739
361	55 751	55 763	55 775	55 787	55 799	55 811	55 823	55 835	55 847	55 859
362	55 871	55 883	55 895	55 907	55 919	55 931	55 943	55 955	55 967	55 979
363	55 991	56 003	56 015	56 027	56 038	56 050	56 062	56 074	56 086	56 098
364	56 110	56 122	56 134	56 146	56 158	56 170	56 182	56 194	56 205	56 217
365	56 229	56 241	56 253	56 265	56 277	56 289	56 301	56 312	56 324	56 336
366	56 348	56 360	56 372	56 384	56 396	56 407	56 419	56 431	56 443	56 455
367	56 467	56 478	56 490	56 502	56 514	56 526	56 538	56 549	56 561	56 573
368	56 585	56 597	56 608	56 620	56 632	56 644	56 656	56 667	56 679	56 691
369	56 703	56 714	56 726	56 738	56 750	56 761	56 773	56 785	56 797	56 808
370	56 820	56 832	56 844	56 855	56 867	56 879	56 891	56 902	56 914	56 926
371	56 937	56 949	56 961	56 972	56 984	56 996	57 008	57 019	57 031	57 043
372	57 054	57 066	57 078	57 089	57 101	57 113	57 124	57 136	57 148	57 159
373	57 171	57 183	57 194	57 206	57 217	57 229	57 241	57 252	57 264	57 276
374	57 287	57 299	57 310	57 322	57 334	57 345	57 357	57 368	57 380	57 392
375	57 403	57 415	57 426	57 438	57 449	57 461	57 473	57 484	57 496	57 507
376	57 519	57 530	57 542	57 553	57 565	57 576	57 588	57 600	57 611	57 623
377	57 634	57 646	57 657	57 669	57 680	57 692	57 703	57 715	57 726	57 738
378	57 749	57 761	57 772	57 784	57 795	57 807	57 818	57 830	57 841	57 852
379	57 864	57 875	57 887	57 898	57 910	57 921	57 933	57 944	57 955	57 967
380	57 978	57 990	58 001	58 013	58 024	58 035	58 047	58 058	58 070	58 081
381	58 092	58 104	58 115	58 127	58 138	58 149	58 161	58 172	58 184	58 195
382	58 206	58 218	58 229	58 240	58 252	58 263	58 274	58 286	58 297	58 309
383	58 320	58 331	58 343	58 354	58 365	58 377	58 388	58 399	58 410	58 422
384	58 433	58 444	58 456	58 467	58 478	58 490	58 501	58 512	58 524	58 535
385	58 546	58 557	58 569	58 580	58 591	58 602	58 614	58 625	58 636	58 647
386	58 659	58 670	58 681	58 692	58 704	58 715	58 726	58 737	58 749	58 760
387	58 771	58 782	58 794	58 805	58 816	58 827	58 838	58 850	58 861	58 872
388	58 883	58 894	58 906	58 917	58 928	58 939	58 950	58 961	58 973	58 984
389	58 995	59 006	59 017	59 028	59 040	59 051	59 062	59 073	59 084	59 095
390	59 106	59 118	59 129	59 140	59 151	59 162	59 173	59 184	59 195	59 207
391	59 218	59 229	59 240	59 251	59 262	59 273	59 284	59 295	59 306	59 318
392	59 329	59 340	59 351	59 362	59 373	59 384	59 395	59 406	59 417	59 428
393	59 439	59 450	59 461	59 472	59 483	59 494	59 506	59 517	59 528	59 539
394	59 550	59 561	59 572	59 583	59 594	59 605	59 616	59 627	59 638	59 649
395	59 660	59 671	59 682	59 693	59 704	59 715	59 726	59 737	59 748	59 759
396	59 770	59 780	59 791	59 802	59 813	59 824	59 835	59 846	59 857	59 868
397	59 879	59 890	59 901	59 912	59 923	59 934	59 945	59 956	59 966	59 977
398	59 988	59 999	60 010	60 021	60 032	60 043	60 054	60 065	60 076	60 086
399	60 097	60 108	60 119	60 130	60 141	60 152	60 163	60 173	60 184	60 195
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—*(Continued)*
400–449

No.	0	1	2	3	4	5	6	7	8	9
400	60 206	60 217	60 228	60 239	60 249	60 260	60 271	60 282	60 293	60 304
401	60 314	60 325	60 336	60 347	60 358	60 369	60 379	60 390	60 401	60 412
402	60 423	60 433	60 444	60 455	60 466	60 477	60 487	60 498	60 509	60 520
403	60 531	60 541	60 552	60 563	60 574	60 584	60 595	60 606	60 617	60 627
404	60 638	60 649	60 660	60 670	60 681	60 692	60 703	60 713	60 724	60 735
405	60 746	60 756	60 767	60 778	60 788	60 799	60 810	60 821	60 831	60 842
406	60 853	60 863	60 874	60 885	60 895	60 906	60 917	60 927	60 938	60 949
407	60 959	60 970	60 981	60 991	61 002	61 013	61 023	61 034	61 045	61 055
408	61 066	61 077	61 087	61 098	61 109	61 119	61 130	61 140	61 151	61 162
409	61 172	61 183	61 194	61 204	61 215	61 225	61 236	61 247	61 257	61 268
410	61 278	61 289	61 300	61 310	61 321	61 331	61 342	61 352	61 363	61 374
411	61 384	61 395	61 405	61 416	61 426	61 437	61 448	61 458	61 469	61 479
412	61 490	61 500	61 511	61 521	61 532	61 542	61 553	61 563	61 574	61 584
413	61 595	61 606	61 616	61 627	61 637	61 648	61 658	61 669	61 679	61 690
414	61 700	61 711	61 721	61 731	61 742	61 752	61 763	61 773	61 784	61 794
415	61 805	61 815	61 826	61 836	61 847	61 857	61 868	61 878	61 888	61 899
416	61 909	61 920	61 930	61 941	61 951	61 962	61 972	61 982	61 993	62 003
417	62 014	62 024	62 034	62 045	62 055	62 066	62 076	62 086	62 097	62 107
418	62 118	62 128	62 138	62 149	62 159	62 170	62 180	62 190	62 201	62 211
419	62 221	62 232	62 242	62 252	62 263	62 273	62 284	62 294	62 304	62 315
420	62 325	62 335	62 346	62 356	62 366	62 377	62 387	62 397	62 408	62 418
421	62 428	62 439	62 449	62 459	62 469	62 480	62 490	62 500	62 511	62 521
422	62 531	62 542	62 552	62 562	62 572	62 583	62 593	62 603	62 613	62 624
423	62 634	62 644	62 655	62 665	62 675	62 685	62 696	62 706	62 716	62 726
424	62 737	62 747	62 757	62 767	62 778	62 788	62 798	62 808	62 818	62 829
425	62 839	62 849	62 859	62 870	62 880	62 890	62 900	62 910	62 921	62 931
426	62 941	62 951	62 961	62 972	62 982	62 992	63 002	63 012	63 022	63 033
427	63 043	63 053	63 063	63 073	63 083	63 094	63 104	63 114	63 124	63 134
428	63 144	63 155	63 165	63 175	63 185	63 195	63 205	63 215	63 225	63 236
429	63 246	63 256	63 266	63 276	63 286	63 296	63 306	63 317	63 327	63 337
430	63 347	63 357	63 367	63 377	63 387	63 397	63 407	63 417	63 428	63 438
431	63 448	63 458	63 468	63 478	63 488	63 498	63 508	63 518	63 528	63 538
432	63 548	63 558	63 568	63 579	63 589	63 599	63 C09	63 619	63 629	63 639
433	63 649	63 659	63 669	63 679	63 689	63 699	63 709	63 719	63 729	63 739
434	63 749	63 759	63 769	63 779	63 789	63 799	63 809	63 819	63 829	63 839
435	63 849	63 859	63 869	63 879	63 889	63 899	63 909	63 919	63 929	63 939
436	63 949	63 959	63 969	63 979	63 988	63 998	64 008	64 018	64 028	64 038
437	64 048	64 058	64 068	64 078	64 088	64 098	64 108	64 118	64 128	64 137
438	64 147	64 157	64 167	64 177	64 187	64 197	64 207	64 217	64 227	64 237
439	64 246	64 256	64 266	64 276	64 286	64 296	64 306	64 316	64 326	64 335
440	64 345	64 355	64 365	64 375	64 385	64 395	64 404	64 414	64 424	64 434
441	64 444	64 454	64 464	64 473	64 483	64 493	64 503	64 513	64 523	64 532
442	64 542	64 552	64 562	64 572	64 582	64 591	64 601	64 611	64 621	64 631
443	64 640	64 650	64 660	64 670	64 680	64 689	64 699	64 709	64 719	64 729
444	64 738	64 748	64 758	64 768	64 777	64 787	64 797	64 807	64 816	64 826
445	64 836	64 846	64 856	64 865	64 875	64 885	64 895	64 904	64 914	64 924
446	64 933	64 943	64 953	64 963	64 972	64 982	64 992	65 002	65 011	65 021
447	65 031	65 040	65 050	65 060	65 070	65 079	65 089	65 099	65 108	65 118
448	65 128	65 137	65 147	65 157	65 167	65 176	65 186	65 196	65 205	65 215
449	65 225	65 234	65 244	65 254	65 263	65 273	65 283	65 292	65 302	65 312
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

450–499

No.	0	1	2	3	4	5	6	7	8	9
450	65 321	65 331	65 341	65 350	65 360	65 369	65 379	65 389	65 398	65 408
451	65 418	65 427	65 437	65 447	65 456	65 466	65 475	65 485	65 495	65 504
452	65 514	65 523	65 533	65 543	65 552	65 562	65 571	65 581	65 591	65 600
453	65 610	65 619	65 629	65 639	65 648	65 658	65 667	65 677	65 686	65 696
454	65 706	65 715	65 725	65 734	65 744	65 753	65 763	65 772	65 782	65 792
455	65 801	65 811	65 820	65 830	65 839	65 849	65 858	65 868	65 877	65 887
456	65 896	65 906	65 916	65 925	65 935	65 944	65 954	65 963	65 973	65 982
457	65 992	66 001	66 011	66 020	66 030	66 039	66 049	66 058	66 068	66 077
458	66 087	66 096	66 106	66 115	66 124	66 134	66 143	66 153	66 162	66 172
459	66 181	66 191	66 200	66 210	66 219	66 229	66 238	66 247	66 257	66 266
460	66 276	66 285	66 295	66 304	66 314	66 323	66 332	66 342	66 351	66 361
461	66 370	66 380	66 389	66 398	66 408	66 417	66 427	66 436	66 445	66 455
462	66 464	66 474	66 483	66 492	66 502	66 511	66 521	66 530	66 539	66 549
463	66 558	66 567	66 577	66 586	66 596	66 605	66 614	66 624	66 633	66 642
464	66 652	66 661	66 671	66 680	66 689	66 699	66 708	66 717	66 727	66 736
465	66 745	66 755	66 764	66 773	66 783	66 792	66 801	66 811	66 820	66 829
466	66 839	66 848	66 857	66 867	66 876	66 885	66 894	66 904	66 913	66 922
467	66 932	66 941	66 950	66 960	66 969	66 978	66 987	66 997	67 006	67 015
468	67 025	67 034	67 043	67 052	67 062	67 071	67 080	67 089	67 099	67 108
469	67 117	67 127	67 136	67 145	67 154	67 164	67 173	67 182	67 191	67 201
470	67 210	67 219	67 228	67 237	67 247	67 256	67 265	67 274	67 284	67 293
471	67 302	67 311	67 321	67 330	67 339	67 348	67 357	67 367	67 376	67 385
472	67 394	67 403	67 413	67 422	67 431	67 440	67 449	67 459	67 468	67 477
473	67 486	67 495	67 504	67 514	67 523	67 532	67 541	67 550	67 560	67 569
474	67 578	67 587	67 596	67 605	67 614	67 624	67 633	67 642	67 651	67 660
475	67 669	67 679	67 688	67 697	67 706	67 715	67 724	67 733	67 742	67 752
476	67 761	67 770	67 779	67 788	67 797	67 806	67 815	67 825	67 834	67 843
477	67 852	67 861	67 870	67 879	67 888	67 897	67 906	67 916	67 925	67 934
478	67 943	67 952	67 961	67 970	67 979	67 988	67 997	68 006	68 015	68 024
479	68 034	68 043	68 052	68 061	68 070	68 079	68 088	68 097	68 106	68 115
480	68 124	68 133	68 142	68 151	68 160	68 169	68 178	68 187	68 196	68 205
481	68 215	68 224	68 233	68 242	68 251	68 260	68 269	68 278	68 287	68 296
482	68 305	68 314	68 323	68 332	68 341	68 350	68 359	68 368	68 377	68 386
483	68 395	68 404	68 413	68 422	68 431	68 440	68 449	68 458	68 467	68 476
484	68 485	68 494	68 502	68 511	68 520	68 529	68 538	68 547	68 556	68 565
485	68 574	68 583	68 592	68 601	68 610	68 619	68 628	68 637	68 646	68 655
486	68 664	68 673	68 681	68 690	68 699	68 708	68 717	68 726	68 735	68 744
487	68 753	68 762	68 771	68 780	68 789	68 797	68 806	68 815	68 824	68 833
488	68 842	68 851	68 860	68 869	68 878	68 886	68 895	68 904	68 913	68 922
489	68 931	68 940	68 949	68 958	68 966	68 975	68 984	68 993	69 002	69 011
490	69 020	69 028	69 037	69 046	69 055	69 064	69 073	69 082	69 090	69 099
491	69 108	69 117	69 126	69 135	69 144	69 152	69 161	69 170	69 179	69 188
492	69 197	69 205	69 214	69 223	69 232	69 241	69 249	69 258	69 267	69 276
493	69 285	69 294	69 302	69 311	69 320	69 329	69 338	69 346	69 355	69 364
494	69 373	69 381	69 390	69 399	69 408	69 417	69 425	69 434	69 443	69 452
495	69 461	69 469	69 478	69 487	69 496	69 504	69 513	69 522	69 531	69 539
496	69 548	69 557	69 566	69 574	69 583	69 592	69 601	69 609	69 618	69 627
497	69 636	69 644	69 653	69 662	69 671	69 679	69 688	69 697	69 705	69 714
498	69 723	69 732	69 740	69 749	69 758	69 767	69 775	69 784	69 793	69 801
499	69 810	69 819	69 827	69 836	69 845	69 854	69 862	69 871	69 880	69 888
No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)
500–549

No.	0	1	2	3	4	5	6	7	8	9
500	69 897	69 906	69 914	69 923	69 932	69 940	69 949	69 958	69 966	69 975
501	69 984	69 992	70 001	70 010	70 018	70 027	70 036	70 044	70 053	70 062
502	70 070	70 079	70 088	70 096	70 105	70 114	70 122	70 131	70 140	70 148
503	70 157	70 165	70 174	70 183	70 191	70 200	70 209	70 217	70 226	70 234
504	70 243	70 252	70 260	70 269	70 278	70 286	70 295	70 303	70 312	70 321
505	70 329	70 338	70 346	70 355	70 364	70 372	70 381	70 389	70 398	70 406
506	70 415	70 424	70 432	70 441	70 449	70 458	70 467	70 475	70 484	70 492
507	70 501	70 509	70 518	70 526	70 535	70 544	70 552	70 561	70 569	70 578
508	70 586	70 595	70 603	70 612	70 621	70 629	70 638	70 646	70 655	70 663
509	70 672	70 680	70 689	70 697	70 706	70 714	70 723	70 731	70 740	70 749
510	70 757	70 766	70 774	70 783	70 791	70 800	70 808	70 817	70 825	70 834
511	70 842	70 851	70 859	70 868	70 876	70 885	70 893	70 902	70 910	70 919
512	70 927	70 935	70 944	70 952	70 961	70 969	70 978	70 986	70 995	71 003
513	71 012	71 020	71 029	71 037	71 046	71 054	71 063	71 071	71 079	71 088
514	71 096	71 105	71 113	71 122	71 130	71 139	71 147	71 155	71 164	71 172
515	71 181	71 189	71 198	71 206	71 214	71 223	71 231	71 240	71 248	71 257
516	71 265	71 273	71 282	71 290	71 299	71 307	71 315	71 324	71 332	71 341
517	71 349	71 357	71 366	71 374	71 383	71 391	71 399	71 408	71 416	71 425
518	71 433	71 441	71 450	71 458	71 466	71 475	71 483	71 492	71 500	71 508
519	71 517	71 525	71 533	71 542	71 550	71 559	71 567	71 575	71 584	71 592
520	71 600	71 609	71 617	71 625	71 634	71 642	71 650	71 659	71 667	71 675
521	71 684	71 692	71 700	71 709	71 717	71 725	71 734	71 742	71 750	71 759
522	71 767	71 775	71 784	71 792	71 800	71 809	71 817	71 825	71 834	71 842
523	71 850	71 858	71 867	71 875	71 883	71 892	71 900	71 908	71 917	71 925
524	71 933	71 941	71 950	71 958	71 966	71 975	71 983	71 991	71 999	72 008
525	72 016	72 024	72 032	72 041	72 049	72 057	72 066	72 074	72 082	72 090
526	72 099	72 107	72 115	72 123	72 132	72 140	72 148	72 156	72 165	72 173
527	72 181	72 189	72 198	72 206	72 214	72 222	72 230	72 239	72 247	72 255
528	72 263	72 272	72 280	72 288	72 296	72 304	72 313	72 321	72 329	72 337
529	72 346	72 354	72 362	72 370	72 378	72 387	72 395	72 403	72 411	72 419
530	72 428	72 436	72 444	72 452	72 460	72 469	72 477	72 485	72 493	72 501
531	72 509	72 518	72 526	72 534	72 542	72 550	72 558	72 567	72 575	72 583
532	72 591	72 599	72 607	72 616	72 624	72 632	72 640	72 648	72 656	72 665
533	72 673	72 681	72 689	72 697	72 705	72 713	72 722	72 730	72 738	72 746
534	72 754	72 762	72 770	72 779	72 787	72 795	72 803	72 811	72 819	72 827
535	72 835	72 843	72 852	72 860	72 868	72 876	72 884	72 892	72 900	72 908
536	72 916	72 925	72 933	72 941	72 949	72 957	72 965	72 973	72 981	72 989
537	72 997	73 006	73 014	73 022	73 030	73 038	73 046	73 054	73 062	73 070
538	73 078	73 086	73 094	73 102	73 111	73 119	73 127	73 135	73 143	73 151
539	73 159	73 167	73 175	73 183	73 191	73 199	73 207	73 215	73 223	73 231
540	73 239	73 247	73 255	73 263	73 272	73 280	73 288	73 296	73 304	73 312
541	73 320	73 328	73 336	73 344	73 352	73 360	73 368	73 376	73 384	73 392
542	73 400	73 408	73 416	73 424	73 432	73 440	73 448	73 456	73 464	73 472
543	73 480	73 488	73 496	73 504	73 512	73 520	73 528	73 536	73 544	73 552
544	73 560	73 568	73 576	73 584	73 592	73 600	73 608	73 616	73 624	73 632
545	73 640	73 648	73 656	73 664	73 672	73 679	73 687	73 695	73 703	73 711
546	73 719	73 727	73 735	73 743	73 751	73 759	73 767	73 775	73 783	73 791
547	73 799	73 807	73 815	73 823	73 830	73 838	73 846	73 854	73 862	73 870
548	73 878	73 886	73 894	73 902	73 910	73 918	73 926	73 933	73 941	73 949
549	73 957	73 965	73 973	73 981	73 989	73 997	74 005	74 013	74 020	74 028
No.	0	1	2	3	4	5	6	7	8	9

Table 7.—Five-place Common Logarithms of Numbers.—(*Continued*)
650–699

No.	0	1	2	3	4	5	6	7	8	9
650	81 291	81 298	81 305	81 311	81 318	81 325	81 331	81 338	81 345	81 351
651	81 358	81 365	81 371	81 378	81 385	81 391	81 398	81 405	81 411	81 418
652	81 425	81 431	81 438	81 445	81 451	81 458	81 465	81 471	81 478	81 485
653	81 491	81 498	81 505	81 511	81 518	81 525	81 531	81 538	81 544	81 551
654	81 558	81 564	81 571	81 578	81 584	81 591	81 598	81 604	81 611	81 617
655	81 624	81 631	81 637	81 644	81 651	81 657	81 664	81 671	81 677	81 684
656	81 690	81 697	81 704	81 710	81 717	81 723	81 730	81 737	81 743	81 750
657	81 757	81 763	81 770	81 776	81 783	81 790	81 796	81 803	81 809	81 816
658	81 823	81 829	81 836	81 842	81 849	81 856	81 862	81 869	81 875	81 882
659	81 889	81 895	81 902	81 908	81 915	81 921	81 928	81 935	81 941	81 948
660	81 954	81 961	81 968	81 974	81 981	81 987	81 994	82 000	82 007	82 014
661	82 020	82 027	82 033	82 040	82 046	82 053	82 060	82 066	82 073	82 079
662	82 086	82 092	82 099	82 105	82 112	82 119	82 125	82 132	82 138	82 145
663	82 151	82 158	82 164	82 171	82 178	82 184	82 191	82 197	82 204	82 210
664	82 217	82 223	82 230	82 236	82 243	82 249	82 256	82 263	82 269	82 276
665	82 282	82 289	82 295	82 302	82 308	82 315	82 321	82 328	82 334	82 341
666	82 347	82 354	82 360	82 367	82 373	82 380	82 387	82 393	82 400	82 406
667	82 413	82 419	82 426	82 432	82 439	82 445	82 452	82 458	82 465	82 471
668	82 478	82 484	82 491	82 497	82 504	82 510	82 517	82 523	82 530	82 536
669	82 543	82 549	82 556	82 562	82 569	82 575	82 582	82 588	82 595	82 601
670	82 607	82 614	82 620	82 627	82 633	82 640	82 646	82 653	82 659	82 666
671	82 672	82 679	82 685	82 692	82 698	82 705	82 711	82 718	82 724	82 730
672	82 737	82 743	82 750	82 756	82 763	82 769	82 776	82 782	82 789	82 795
673	82 802	82 808	82 814	82 821	82 827	82 834	82 840	82 847	82 853	82 860
674	82 866	82 872	82 879	82 885	82 892	82 898	82 905	82 911	82 918	82 924
675	82 930	82 937	82 943	82 950	82 956	82 963	82 969	82 975	82 982	82 988
676	82 995	83 001	83 008	83 014	83 020	83 027	83 033	83 040	83 046	83 052
677	83 059	83 065	83 072	83 078	83 085	83 091	83 097	83 104	83 110	83 117
678	83 123	83 129	83 136	83 142	83 149	83 155	83 161	83 168	83 174	83 181
679	83 187	83 193	83 200	83 206	83 213	83 219	83 225	83 232	83 238	83 245
680	83 251	83 257	83 264	83 270	83 276	83 283	83 289	83 296	83 302	83 308
681	83 315	83 321	83 327	83 334	83 340	83 347	83 353	83 359	83 366	83 372
682	83 378	83 385	83 391	83 398	83 404	83 410	83 417	83 423	83 429	83 436
683	83 442	83 448	83 455	83 461	83 467	83 474	83 480	83 487	83 493	83 499
684	83 506	83 512	83 518	83 525	83 531	83 537	83 544	83 550	83 556	83 563
685	83 569	83 575	83 582	83 588	83 594	83 601	83 607	83 613	83 620	83 626
686	83 632	83 639	83 645	83 651	83 658	83 664	83 670	83 677	83 683	83 689
687	83 696	83 702	83 708	83 715	83 721	83 727	83 734	83 740	83 746	83 753
688	83 759	83 765	83 771	83 778	83 784	83 790	83 797	83 803	83 809	83 816
689	83 822	83 828	83 835	83 841	83 847	83 853	83 860	83 866	83 872	83 879
690	83 885	83 891	83 897	83 904	83 910	83 916	83 923	83 929	83 935	83 942
691	83 948	83 954	83 960	83 967	83 973	83 979	83 985	83 992	83 998	84 004
692	84 011	84 017	84 023	84 029	84 036	84 042	84 048	84 055	84 061	84 067
693	84 073	84 080	84 086	84 092	84 098	84 105	84 111	84 117	84 123	84 130
694	84 136	84 142	84 148	84 155	84 161	84 167	84 173	84 180	84 186	84 192
695	84 198	84 205	84 211	84 217	84 223	84 230	84 236	84 242	84 248	84 255
696	84 261	84 267	84 273	84 280	84 286	84 292	84 298	84 305	84 311	84 317
697	84 323	84 330	84 336	84 342	84 348	84 354	84 361	84 367	84 373	84 379
698	84 386	84 392	84 398	84 404	84 410	84 417	84 423	84 429	84 435	84 442
699	84 448	84 454	84 460	84 466	84 473	84 479	84 485	84 491	84 497	84 504
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

700–749

No.	0	1	2	3	4	5	6	7	8	9
700	84 510	84 516	84 522	84 528	84 535	84 541	84 547	84 553	84 559	84 566
701	84 572	84 578	84 584	84 590	84 597	84 603	84 609	84 615	84 621	84 628
702	84 634	84 640	84 646	84 652	84 658	84 665	84 671	84 677	84 683	84 689
703	84 696	84 702	84 708	84 714	84 720	84 726	84 733	84 739	84 745	84 751
704	84 757	84 763	84 770	84 776	84 782	84 788	84 794	84 800	84 807	84 813
705	84 819	84 825	84 831	84 837	84 844	84 850	84 856	84 862	84 868	84 874
706	84 880	84 887	84 893	84 899	84 905	84 911	84 917	84 924	84 930	84 936
707	84 942	84 948	84 954	84 960	84 967	84 973	84 979	84 985	84 991	84 997
708	85 003	85 009	85 016	85 022	85 028	85 034	85 040	85 046	85 052	85 058
709	85 065	85 071	85 077	85 083	85 089	85 095	85 101	85 107	85 114	85 120
710	85 126	85 132	85 138	85 144	85 150	85 156	85 163	85 169	85 175	85 181
711	85 187	85 193	85 199	85 205	85 211	85 217	85 224	85 230	85 236	85 242
712	85 248	85 254	85 260	85 266	85 272	85 278	85 285	85 291	85 297	85 303
713	85 309	85 315	85 321	85 327	85 333	85 339	85 345	85 352	85 358	85 364
714	85 370	85 376	85 382	85 388	85 394	85 400	85 406	85 412	85 418	85 425
715	85 431	85 437	85 443	85 449	85 455	85 461	85 467	85 473	85 479	85 485
716	85 491	85 497	85 503	85 509	85 516	85 522	85 528	85 534	85 540	85 546
717	85 552	85 558	85 564	85 570	85 576	85 582	85 588	85 594	85 600	85 606
718	85 612	85 618	85 625	85 631	85 637	85 643	85 649	85 655	85 661	85 667
719	85 673	85 679	85 685	85 691	85 697	85 703	85 709	85 715	85 721	85 727
720	85 733	85 739	85 745	85 751	85 757	85 763	85 769	85 775	85 781	85 788
721	85 794	85 800	85 806	85 812	85 818	85 824	85 830	85 836	85 842	85 848
722	85 854	85 860	85 866	85 872	85 878	85 884	85 890	85 896	85 902	85 908
723	85 914	85 920	85 926	85 932	85 938	85 944	85 950	85 956	85 962	85 968
724	85 974	85 980	85 986	85 992	85 998	86 004	86 010	86 016	86 022	86 028
725	86 034	86 040	86 046	86 052	86 058	86 064	86 070	86 076	86 082	86 088
726	86 094	86 100	86 106	86 112	86 118	86 124	86 130	86 136	86 141	86 147
727	86 153	86 159	86 165	86 171	86 177	86 183	86 189	86 195	86 201	86 207
728	86 213	86 219	86 225	86 231	86 237	86 243	86 249	86 255	86 261	86 267
729	86 273	86 279	86 285	86 291	86 297	86 303	86 308	86 314	86 320	86 326
730	86 332	86 338	86 344	86 350	86 356	86 362	86 368	86 374	86 380	86 386
731	86 392	86 398	86 404	86 410	86 415	86 421	86 427	86 433	86 439	86 445
732	86 451	86 457	86 463	86 469	86 475	86 481	86 487	86 493	86 499	86 504
733	86 510	86 516	86 522	86 528	86 534	86 540	86 546	86 552	86 558	86 564
734	86 570	86 576	86 581	86 587	86 593	86 599	86 605	86 611	86 617	86 623
735	86 629	86 635	86 641	86 646	86 652	86 658	86 664	86 670	86 676	86 682
736	86 688	86 694	86 700	86 705	86 711	86 717	86 723	86 729	86 735	86 741
737	86 747	86 753	86 759	86 764	86 770	86 776	86 782	86 788	86 794	86 800
738	86 806	86 812	86 817	86 823	86 829	86 835	86 841	86 847	86 853	86 859
739	86 864	86 870	86 876	86 882	86 888	86 894	86 900	86 906	86 911	86 917
740	86 923	86 929	86 935	86 941	86 947	86 953	86 958	86 964	86 970	86 976
741	86 982	86 988	86 994	86 999	87 005	87 011	87 017	87 023	87 029	87 035
742	87 040	87 046	87 052	87 058	87 064	87 070	87 075	87 081	87 087	87 093
743	87 099	87 105	87 111	87 116	87 122	87 128	87 134	87 140	87 146	87 151
744	87 157	87 163	87 169	87 175	87 181	87 186	87 192	87 198	87 204	87 210
745	87 216	87 221	87 227	87 233	87 239	87 245	87 251	87 256	87 262	87 268
746	87 274	87 280	87 286	87 291	87 297	87 303	87 309	87 315	87 320	87 326
747	87 332	87 338	87 344	87 349	87 355	87 361	87 367	87 373	87 379	87 384
748	87 390	87 396	87 402	87 408	87 413	87 419	87 425	87 431	87 437	87 442
749	87 448	87 454	87 460	87 466	87 471	87 477	87 483	87 489	87 495	87 500

No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)
750–799

No.	0	1	2	3	4	5	6	7	8	9
750	87 506	87 512	87 518	87 523	87 529	87 535	87 541	87 547	87 552	87 558
751	87 564	87 570	87 576	87 581	87 587	87 593	87 599	87 604	87 610	87 616
752	87 622	87 628	87 633	87 639	87 645	87 651	87 656	87 662	87 668	87 674
753	87 679	87 685	87 691	87 697	87 703	87 708	87 714	87 720	87 726	87 731
754	87 737	87 743	87 749	87 754	87 760	87 766	87 772	87 777	87 783	87 789
755	87 795	87 800	87 806	87 812	87 818	87 823	87 829	87 835	87 841	87 846
756	87 852	87 858	87 864	87 869	87 875	87 881	87 887	87 892	87 898	87 904
757	87 910	87 915	87 921	87 927	87 933	87 938	87 944	87 950	87 955	87 961
758	87 967	87 973	87 978	87 984	87 990	87 996	88 001	88 007	88 013	88 018
759	88 024	88 030	88 036	88 041	88 047	88 053	88 058	88 064	88 070	88 076
760	88 081	88 087	88 093	88 098	88 104	88 110	88 116	88 121	88 127	88 133
761	88 138	88 144	88 150	88 156	88 161	88 167	88 173	88 178	88 184	88 190
762	88 195	88 201	88 207	88 213	88 218	88 224	88 230	88 235	88 241	88 247
763	88 252	88 258	88 264	88 270	88 275	88 281	88 287	88 292	88 298	88 304
764	88 309	88 315	88 321	88 326	88 332	88 338	88 343	88 349	88 355	88 360
765	88 366	88 372	88 377	88 383	88 389	88 395	88 400	88 406	88 412	88 417
766	88 423	88 429	88 434	88 440	88 446	88 451	88 457	88 463	88 468	88 474
767	88 480	88 485	88 491	88 497	88 502	88 508	88 513	88 519	88 525	88 530
768	88 536	88 542	88 547	88 553	88 559	88 564	88 570	88 576	88 581	88 587
769	88 593	88 598	88 604	88 610	88 615	88 621	88 627	88 632	88 638	88 643
770	88 649	88 655	88 660	88 666	88 672	88 677	88 683	88 689	88 694	88 700
771	88 705	88 711	88 717	88 722	88 728	88 734	88 739	88 745	88 750	88 756
772	88 762	88 767	88 773	88 779	88 784	88 790	88 795	88 801	88 807	88 812
773	88 818	88 824	88 829	88 835	88 840	88 846	88 852	88 857	88 863	88 868
774	88 874	88 880	88 885	88 891	88 897	88 902	88 908	88 913	88 919	88 925
775	88 930	88 936	88 941	88 947	88 953	88 958	88 964	88 969	88 975	88 981
776	88 986	88 992	88 997	89 003	89 009	89 014	89 020	89 025	89 031	89 037
777	89 042	89 048	89 053	89 059	89 064	89 070	89 076	89 081	89 087	89 092
778	89 098	89 104	89 109	89 115	89 120	89 126	89 131	89 137	89 143	89 148
779	89 154	89 159	89 165	89 170	89 176	89 182	89 187	89 193	89 198	89 204
780	89 209	89 215	89 221	89 226	89 232	89 237	89 243	89 248	89 254	89 260
781	89 265	89 271	89 276	89 282	89 287	89 293	89 298	89 304	89 310	89 315
782	89 321	89 326	89 332	89 337	89 343	89 348	89 354	89 360	89 365	89 371
783	89 376	89 382	89 387	89 393	89 398	89 404	89 409	89 415	89 421	89 426
784	89 432	89 437	89 443	89 448	89 454	89 459	89 465	89 470	89 476	89 481
785	89 487	89 492	89 498	89 504	89 509	89 515	89 520	89 526	89 531	89 537
786	89 542	89 548	89 553	89 559	89 564	89 570	89 575	89 581	89 586	89 592
787	89 597	89 603	89 609	89 614	89 620	89 625	89 631	89 636	89 642	89 647
788	89 653	89 658	89 664	89 669	89 675	89 680	89 686	89 691	89 697	89 702
789	89 708	89 713	89 719	89 724	89 730	89 735	89 741	89 746	89 752	89 757
790	89 763	89 768	89 774	89 779	89 785	89 790	89 796	89 801	89 807	89 812
791	89 818	89 823	89 829	89 834	89 840	89 845	89 851	89 856	89 862	89 867
792	89 873	89 878	89 883	89 889	89 894	89 900	89 905	89 911	89 916	89 922
793	89 927	89 933	89 938	89 944	89 949	89 955	89 960	89 966	89 971	89 977
794	89 982	89 988	89 993	89 998	90 004	90 009	90 015	90 020	90 026	90 031
795	90 037	90 042	90 048	90 053	90 059	90 064	90 069	90 075	90 080	90 086
796	90 091	90 097	90 102	90 108	90 113	90 119	90 124	90 129	90 135	90 140
797	90 146	90 151	90 157	90 162	90 168	90 173	90 179	90 184	90 189	90 195
798	90 200	90 206	90 211	90 217	90 222	90 227	90 233	90 238	90 244	90 249
799	90 255	90 260	90 266	90 271	90 276	90 282	90 287	90 293	90 298	90 304
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

800–849

No.	0	1	2	3	4	5	6	7	8	9
800	90 309	90 314	90 320	90 325	90 331	90 336	90 342	90 347	90 352	90 358
801	90 363	90 369	90 374	90 380	90 385	90 390	90 396	90 401	90 407	90 412
802	90 417	90 423	90 428	90 434	90 439	90 445	90 450	90 455	90 461	90 466
803	90 472	90 477	90 482	90 488	90 493	90 499	90 504	90 509	90 515	90 520
804	90 526	90 531	90 536	90 542	90 547	90 553	90 558	90 563	90 569	90 574
805	90 580	90 585	90 590	90 596	90 601	90 607	90 612	90 617	90 623	90 628
806	90 634	90 639	90 644	90 650	90 655	90 660	90 666	90 671	90 677	90 682
807	90 687	90 693	90 698	90 703	90 709	90 714	90 720	90 725	90 730	90 736
808	90 741	90 747	90 752	90 757	90 763	90 768	90 773	90 779	90 784	90 789
809	90 795	90 800	90 806	90 811	90 816	90 822	90 827	90 832	90 838	90 843
810	90 849	90 854	90 859	90 865	90 870	90 875	90 881	90 886	90 891	90 897
811	90 902	90 907	90 913	90 918	90 924	90 929	90 934	90 940	90 945	90 950
812	90 956	90 961	90 966	90 972	90 977	90 982	90 988	90 993	90 998	91 004
813	91 009	91 014	91 020	91 025	91 030	91 036	91 041	91 046	91 052	91 057
814	91 062	91 068	91 073	91 078	91 084	91 089	91 094	91 100	91 105	91 110
815	91 116	91 121	91 126	91 132	91 137	91 142	91 148	91 153	91 158	91 164
816	91 169	91 174	91 180	91 185	91 190	91 196	91 201	91 206	91 212	91 217
817	91 222	91 228	91 233	91 238	91 243	91 249	91 254	91 259	91 265	91 270
818	91 275	91 281	91 286	91 291	91 297	91 302	91 307	91 312	91 318	91 323
819	91 328	91 334	91 339	91 344	91 350	91 355	91 360	91 365	91 371	91 376
820	91 381	91 387	91 392	91 397	91 403	91 408	91 413	91 418	91 424	91 429
821	91 434	91 440	91 445	91 450	91 455	91 461	91 466	91 471	91 477	91 482
822	91 487	91 492	91 498	91 503	91 508	91 514	91 519	91 524	91 529	91 535
823	91 540	91 545	91 551	91 556	91 561	91 566	91 572	91 577	91 582	91 587
824	91 593	91 598	91 603	91 609	91 614	91 619	91 624	91 630	91 635	91 640
825	91 645	91 651	91 656	91 661	91 666	91 672	91 677	91 682	91 687	91 693
826	91 698	91 703	91 709	91 714	91 719	91 724	91 730	91 735	91 740	91 745
827	91 751	91 756	91 761	91 766	91 772	91 777	91 782	91 787	91 793	91 798
828	91 803	91 808	91 814	91 819	91 824	91 829	91 834	91 840	91 845	91 850
829	91 855	91 861	91 866	91 871	91 876	91 882	91 887	91 892	91 897	91 903
830	91 908	91 913	91 918	91 924	91 929	91 934	91 939	91 944	91 950	91 955
831	91 960	91 965	91 971	91 976	91 981	91 986	91 991	91 997	92 002	92 007
832	92 012	92 018	92 023	92 028	92 033	92 038	92 044	92 049	92 054	92 059
833	92 065	92 070	92 075	92 080	92 085	92 091	92 096	92 101	92 106	92 111
834	92 117	92 122	92 127	92 132	92 137	92 143	92 148	92 153	92 158	92 163
835	92 169	92 174	92 179	92 184	92 189	92 195	92 200	92 205	92 210	92 215
836	92 221	92 226	92 231	92 236	92 241	92 247	92 252	92 257	92 262	92 267
837	92 273	92 278	92 283	92 288	92 293	92 298	92 304	92 309	92 314	92 319
838	92 324	92 330	92 335	92 340	92 345	92 350	92 355	92 361	92 366	92 371
839	92 376	92 381	92 387	92 392	92 397	92 402	92 407	92 412	92 418	92 423
840	92 428	92 433	92 438	92 443	92 449	92 454	92 459	92 464	92 469	92 474
841	92 480	92 485	92 490	92 495	92 500	92 505	92 511	92 516	92 521	92 526
842	92 531	92 536	92 542	92 547	92 552	92 557	92 562	92 567	92 572	92 578
843	92 583	92 588	92 593	92 598	92 603	92 609	92 614	92 619	92 624	92 629
844	92 634	92 639	92 645	92 650	92 655	92 660	92 665	92 670	92 675	92 681
845	92 686	92 691	92 696	92 701	92 706	92 711	92 716	92 722	92 727	92 732
846	92 737	92 742	92 747	92 752	92 758	92 763	92 768	92 773	92 778	92 783
847	92 788	92 793	92 799	92 804	92 809	92 814	92 819	92 824	92 829	92 834
848	92 840	92 845	92 850	92 855	92 860	92 865	92 870	92 875	92 881	92 886
849	92 891	92 896	92 901	92 906	92 911	92 916	92 921	92 927	92 932	92 937
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(Continued)
850–899

No.	0	1	2	3	4	5	6	7	8	9
850	92 942	92 947	92 952	92 957	92 962	92 967	92 973	92 978	92 983	92 988
851	92 993	92 998	93 003	93 008	93 013	93 018	93 024	93 029	93 034	93 039
852	93 044	93 049	93 054	93 059	93 064	93 069	93 075	93 080	93 085	93 090
853	93 095	93 100	93 105	93 110	93 115	93 120	93 125	93 131	93 136	93 141
854	93 146	93 151	93 156	93 161	93 166	93 171	93 176	93 181	93 186	93 192
855	93 197	93 202	93 207	93 212	93 217	93 222	93 227	93 232	93 237	93 242
856	93 247	93 252	93 258	93 263	93 268	93 273	93 278	93 283	93 288	93 293
857	93 298	93 303	93 308	93 313	93 318	93 323	93 328	93 334	93 339	93 344
858	93 349	93 354	93 359	93 364	93 369	93 374	93 379	93 384	93 389	93 394
859	93 399	93 404	93 409	93 414	93 420	93 425	93 430	93 435	93 440	93 445
860	93 450	93 455	93 460	93 465	93 470	93 475	93 480	93 485	93 490	93 495
861	93 500	93 505	93 510	94 515	93 520	93 526	93 531	93 536	93 541	93 546
862	93 551	93 556	93 561	93 566	93 571	93 576	93 581	93 586	93 591	93 596
863	93 601	93 606	93 611	93 616	93 621	93 626	93 631	93 636	93 641	93 646
864	93 651	93 656	93 661	93 666	93 671	93 676	93 682	93 687	93 692	93 697
865	93 702	93 707	93 712	93 717	93 722	93 727	93 732	93 737	93 742	93 747
866	93 752	93 757	93 762	93 767	93 772	93 777	93 782	93 787	93 792	93 797
867	93 802	93 807	93 812	93 817	93 822	93 827	93 832	93 837	93 842	93 847
868	93 852	93 857	93 862	93 867	93 872	93 877	93 882	93 887	93 892	93 897
869	93 902	93 907	93 912	93 917	93 922	93 927	93 932	93 937	93 942	93 947
870	93 952	93 957	93 962	93 967	93 972	93 977	93 982	93 987	93 992	93 997
871	94 002	94 007	94 012	94 017	94 022	94 027	94 032	94 037	94 042	94 047
872	94 052	94 057	94 062	94 067	94 072	94 077	94 082	94 086	94 091	94 096
873	94 101	94 106	94 111	94 116	94 121	94 126	94 131	94 136	94 141	94 146
874	94 151	94 156	94 161	94 166	94 171	94 176	94 181	94 186	94 191	94 196
875	94 201	94 206	94 211	94 216	94 221	94 226	94 231	94 236	94 240	94 245
876	94 250	94 255	94 260	94 265	94 270	94 275	94 280	94 285	94 290	94 295
877	94 300	94 305	94 310	94 315	94 320	94 325	94 330	94 335	94 340	94 345
878	94 349	94 354	94 359	94 364	94 369	94 374	94 379	94 384	94 389	94 394
879	94 399	94 404	94 409	94 414	94 419	94 424	94 429	94 433	94 438	94 443
880	94 448	94 453	94 458	94 463	94 468	94 473	94 478	94 483	94 488	94 493
881	94 498	94 503	94 507	94 512	94 517	94 522	94 527	94 532	94 537	94 542
882	94 547	94 552	94 557	94 562	94 567	94 571	94 576	94 581	94 586	94 591
883	94 596	94 601	94 606	94 611	94 616	94 621	94 626	94 630	94 635	94 640
884	94 645	94 650	94 655	94 660	94 665	94 670	94 675	94 680	94 685	94 689
885	94 694	94 699	94 704	94 709	94 714	94 719	94 724	94 729	94 734	94 738
886	94 743	94 748	94 753	94 758	94 763	94 768	94 773	94 778	94 783	94 787
887	94 792	94 797	94 802	94 807	94 812	94 817	94 822	94 827	94 832	94 836
888	94 841	94 846	94 851	94 856	94 861	94 866	94 871	94 876	94 880	94 885
889	94 890	94 895	94 900	94 905	94 910	94 915	94 919	94 924	94 929	94 934
890	94 939	94 944	94 949	94 954	94 959	94 963	94 968	94 973	94 978	94 983
891	94 988	94 993	94 998	95 002	95 007	95 012	95 017	95 022	95 027	95 032
892	95 036	95 041	95 046	95 051	95 056	95 061	95 066	95 071	95 075	95 080
893	95 085	95 090	95 095	95 100	95 105	95 109	95 114	95 119	95 124	95 129
894	95 134	95 139	95 143	95 148	95 153	95 158	95 163	95 168	95 173	95 177
895	95 182	95 187	95 192	95 197	95 202	95 207	95 211	95 216	95 221	95 226
896	95 231	95 236	95 240	95 245	95 250	95 255	95 260	95 265	95 270	95 274
897	95 279	95 284	95 289	95 294	95 299	95 303	95 308	95 313	95 318	95 323
898	95 328	95 332	95 337	95 342	95 347	95 352	95 357	95 361	95 366	95 371
899	95 376	95 381	95 386	95 390	95 395	95 400	95 405	95 410	95 415	95 419
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—*(Continued)*
900–949

No.	0	1	2	3	4	5	6	7	8	9
900	95 424	95 429	95 434	95 439	95 444	95 448	95 453	95 458	95 463	95 468
901	95 472	95 477	95 482	95 487	95 492	95 497	95 501	95 506	95 511	95 516
902	95 521	95 525	95 530	95 535	95 540	95 545	95 550	95 554	95 559	95 564
903	95 569	95 574	95 578	95 583	95 588	95 593	95 598	95 602	95 607	95 612
904	95 617	95 622	95 626	95 631	95 636	95 641	95 646	95 650	95 655	95 660
905	95 665	95 670	95 674	95 679	95 684	95 689	95 694	95 698	95 703	95 708
906	95 713	95 718	95 722	95 727	95 732	95 737	95 742	95 746	95 751	95 756
907	95 761	95 766	95 770	95 775	95 780	95 785	95 789	95 794	95 799	95 804
908	95 809	95 813	95 818	95 823	95 828	95 832	95 837	95 842	95 847	95 852
909	95 856	95 861	95 866	95 871	95 875	95 880	95 885	95 890	95 895	95 899
910	95 904	95 909	95 914	95 918	95 923	95 928	95 933	95 938	95 942	95 947
911	95 952	95 957	95 961	95 966	95 971	95 976	95 980	95 985	95 990	95 995
912	95 999	96 004	96 009	96 014	96 019	96 023	96 028	96 033	96 038	96 042
913	96 047	96 052	96 057	96 061	96 066	96 071	96 076	96 080	96 085	96 090
914	96 095	96 099	96 104	96 109	96 114	96 118	96 123	96 128	96 133	96 137
915	96 142	96 147	96 152	96 156	96 161	96 166	96 171	96 175	96 180	96 185
916	96 190	96 194	96 199	96 204	96 209	96 213	96 218	96 223	96 227	96 232
917	96 237	96 242	96 246	96 251	96 256	96 261	96 265	96 270	96 275	96 280
918	96 284	96 289	96 294	96 298	96 303	96 308	96 313	96 317	96 322	96 327
919	96 332	96 336	96 341	96 346	96 350	96 355	96 360	96 365	96 369	96 374
920	96 379	96 384	96 388	96 393	96 398	96 402	96 407	96 412	96 417	96 421
921	96 426	96 431	96 435	96 440	96 445	96 450	96 454	96 459	96 464	96 468
922	96 473	96 478	96 483	96 487	96 492	96 497	96 501	96 506	96 511	96 515
923	96 520	96 525	96 530	96 534	96 539	96 544	96 548	96 553	96 558	96 562
924	96 567	96 572	96 577	96 581	96 586	96 591	96 595	96 600	96 605	96 609
925	96 614	96 619	96 624	96 628	96 633	96 638	96 642	96 647	96 652	96 656
926	96 661	96 666	96 670	96 675	96 680	96 685	96 689	96 694	96 699	96 703
927	96 708	96 713	96 717	96 722	96 727	96 731	96 736	96 741	96 745	96 750
928	96 755	96 759	96 764	96 769	96 774	96 778	96 783	96 788	96 792	96 797
929	96 802	96 806	96 811	96 816	96 820	96 825	96 830	96 834	96 839	96 844
930	96 848	96 853	96 858	96 862	96 867	96 872	96 876	96 881	96 886	96 890
931	96 895	96 900	96 904	96 909	96 914	96 918	96 923	96 928	96 932	96 937
932	96 942	96 946	96 951	96 956	96 960	96 965	96 970	96 974	96 979	96 984
933	96 988	96 993	96 997	97 002	97 007	97 011	97 016	97 021	97 025	97 030
934	97 035	97 039	97 044	97 049	97 053	97 058	97 063	97 067	97 072	97 077
935	97 081	97 086	97 090	97 095	97 100	97 104	97 109	97 114	97 118	97 123
936	97 128	97 132	97 137	97 142	97 146	97 151	97 155	97 160	97 165	97 169
937	97 174	97 179	97 183	97 188	97 192	97 197	97 202	97 206	97 211	97 216
938	97 220	97 225	97 230	97 234	97 239	97 243	97 248	97 253	97 257	97 262
939	97 267	97 271	97 276	97 280	97 285	97 290	97 294	97 299	97 304	97 308
940	97 313	97 317	97 322	97 327	97 331	97 336	97 340	97 345	97 350	97 354
941	97 359	97 364	97 368	97 373	97 377	97 382	97 387	97 391	97 396	97 400
942	97 405	97 410	97 414	97 419	97 424	97 428	97 433	97 437	97 442	97 447
943	97 451	97 456	97 460	97 465	97 470	97 474	97 479	97 483	97 488	97 493
944	97 497	97 502	97 506	97 511	97 516	97 520	97 525	97 529	97 534	97 539
945	97 543	97 548	97 552	97 557	97 562	97 566	97 571	97 575	97 580	97 585
946	97 589	97 594	97 598	97 603	97 607	97 612	97 617	97 621	97 626	97 630
947	97 635	97 640	97 644	97 649	97 653	97 658	97 663	97 667	97 672	97 676
948	97 681	97 685	97 690	97 695	97 699	97 704	97 708	97 713	97 717	97 722
949	97 727	97 731	97 736	97 740	97 745	97 749	97 754	97 759	97 763	97 768
No.	0	1	2	3	4	5	6	7	8	9

TABLE 7.—FIVE-PLACE COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

950–1000

No.	0	1	2	3	4	5	6	7	8	9
950	97 772	97 777	97 782	97 786	97 791	97 795	97 800	97 804	97 809	97 813
951	97 818	97 823	97 827	97 832	97 836	97 841	97 845	97 850	97 855	97 859
952	97 864	97 868	97 873	97 877	97 882	97 886	97 891	97 896	97 900	97 905
953	97 909	97 914	97 918	97 923	97 928	97 932	97 937	97 941	97 946	97 950
954	97 955	97 959	97 964	97 968	97 973	97 978	97 982	97 987	97 991	97 996
955	98 000	98 005	98 009	98 014	98 019	98 023	98 028	98 032	98 037	98 041
956	98 046	98 050	98 055	98 059	98 064	98 068	98 073	98 078	98 082	98 087
957	98 091	98 096	98 100	98 105	98 109	98 114	98 118	98 123	98 127	98 132
958	98 137	98 141	98 146	98 150	98 155	98 159	98 164	98 168	98 173	98 177
959	98 182	98 186	98 191	98 195	98 200	98 204	98 209	98 214	98 218	98 223
960	98 227	98 232	98 236	98 241	98 245	98 250	98 254	98 259	98 263	98 268
961	98 272	98 277	98 281	98 286	98 290	98 295	98 299	98 304	98 308	98 313
962	98 318	98 322	98 327	98 331	98 336	98 340	98 345	98 349	98 354	98 358
963	98 363	98 367	98 372	98 376	98 381	98 385	98 390	98 394	98 399	98 403
964	98 408	98 412	98 417	98 421	98 426	98 430	98 435	98 439	98 444	98 448
965	98 453	98 457	98 462	98 466	98 471	98 475	98 480	98 484	98 489	98 493
966	98 498	98 502	98 507	98 511	98 516	98 520	98 525	98 529	98 534	98 538
967	98 543	98 547	98 552	98 556	98 561	98 565	98 570	98 574	98 579	98 583
968	98 588	98 592	98 597	98 601	98 605	98 610	98 614	98 619	98 623	98 628
969	98 632	98 637	98 641	98 646	98 650	98 655	98 659	98 664	98 668	98 673
970	98 677	98 682	98 686	98 691	98 695	98 700	98 704	98 709	98 713	98 717
971	98 722	98 726	98 731	98 735	98 740	98 744	98 749	98 753	98 758	98 762
972	98 767	98 771	98 776	98 780	98 784	98 789	98 793	98 798	98 802	98 807
973	98 811	98 816	98 820	98 825	98 829	98 834	98 838	98 843	98 847	98 851
974	98 856	98 860	98 865	98 869	98 874	98 878	98 883	98 887	98 892	98 896
975	98 900	98 905	98 909	98 914	98 918	98 923	98 927	98 932	98 936	98 941
976	98 945	98 949	98 954	98 958	98 963	98 967	98 972	98 976	98 981	98 985
977	98 989	98 994	98 998	99 003	99 007	99 012	99 016	99 021	99 025	99 029
978	99 034	99 038	99 043	99 047	99 052	99 056	99 061	99 065	99 069	99 074
979	99 078	99 083	99 087	99 092	99 096	99 100	99 105	99 109	99 114	99 118
980	99 123	99 127	99 131	99 136	99 140	99 145	99 149	99 154	99 158	99 162
981	99 167	99 171	99 176	99 180	99 185	99 189	99 193	99 198	99 202	99 207
982	99 211	99 216	99 220	99 224	99 229	99 233	99 238	99 242	99 247	99 251
983	99 255	99 260	99 264	99 269	99 273	99 277	99 282	99 286	99 291	99 295
984	99 300	99 304	99 308	99 313	99 317	99 322	99 326	99 330	99 335	99 339
985	99 344	99 348	99 352	99 357	99 361	99 366	99 370	99 374	99 379	99 383
986	99 388	99 392	99 396	99 401	99 405	99 410	99 414	99 419	99 423	99 427
987	99 432	99 436	99 441	99 445	99 449	99 454	99 458	99 463	99 467	99 471
988	99 476	99 480	99 484	99 489	99 493	99 498	99 502	99 506	99 511	99 515
989	99 520	99 524	99 528	99 533	99 537	99 542	99 546	99 550	99 555	99 559
990	99 564	99 568	99 572	99 577	99 581	99 585	99 590	99 594	99 599	99 603
991	99 607	99 612	99 616	99 621	99 625	99 629	99 634	99 638	99 642	99 647
992	99 651	99 656	99 660	99 664	99 669	99 673	99 677	99 682	99 686	99 691
993	99 695	99 699	99 704	99 708	99 712	99 717	99 721	99 726	99 730	99 734
994	99 739	99 743	99 747	99 752	99 756	99 760	99 765	99 769	99 774	99 778
995	99 782	99 787	99 791	99 795	99 800	99 804	99 808	99 813	99 817	99 822
996	99 826	99 830	99 835	99 839	99 843	99 848	99 852	99 856	99 861	99 865
997	99 870	99 874	99 878	99 883	99 887	99 891	99 896	99 900	99 904	99 909
998	99 913	99 917	99 922	99 926	99 930	99 935	99 939	99 944	99 948	99 952
999	99 957	99 961	99 965	99 970	99 974	99 978	99 983	99 987	99 991	99 996
1000	00 000	00 004	00 009	00 013	00 017	00 022	00 026	00 030	00 035	00 039
No.	0	1	2	3	4	5	6	7	8	9

Index